

Joshua Ainsley, PhD
Postdoctoral Researcher
Lab of Leon Reijmers
Neuroscience Department
Tufts University
joshua.ainsley@tufts.edu

RNA-Seq

Lecture Outline

- Acknowledgements
- Computational Biology Initiative
- Course goals/outline
- Cluster test
- Sequencing technologies
- Planning an RNA-seq experiment
- Library prep
- Design exercises
- Some helpful resources

Acknowledgements & Sources

Wash U Genome Center

Vince Magrini, Jason Walker, Sean McGrath

Broad Institute

Mike Zody, Jim Robinson

European Molecular Biology Laboratory

Wolfgang Huber, Simon Anders

Michigan State University

C. Titus Brown

University of Virginia

Aaron Quinlan

Tufts University

Lax Iyer, Peter Castaldi, CBI

CBI: What do we do?

- The Computational Biology Initiative (CBI) is a forum for Tufts researchers to collaborate and develop competitive grants in ‘omics research.
- Beta Site: sites.tufts.edu/cbi
- Please send us your comments about what you would like to see here!

Our Partners

Tufts Medical Center

 Tufts University
Core Facility

C N R
Tufts Center for Neuroscience Research

Tufts | CTSI

uit University Information Technology
Tufts Clinical and Translational Science Institute

Provost's office: Tufts Collaborates! and Tufts Innovates! Grant Awards
Kirby Johnson @Vice Provost Office

 **CBI** Computational Biology Initiative

Specific Aims

- **Establish**
 - A core group of CBI experts
 - Computational Biology resources
 - dynamic web presence and computing resources
- **Raise awareness and educate researchers in genomics**

How can we work together?

- Discuss your research projects
- Ideas for Symposium/talks/courses
- Contribute to our website: Your ideas, protocols/how to
- Keeping up to date with developments
- **Developing new scientific collaborations across disciplines!**

More about CBI

- Attend an open meeting of CBI to discuss how we could work together and what needs to be done!
- Easiest way to contact: lax.iyer@tufts.edu
 - Peter Castaldi, TMC
 - Larry Parnell, HNRCA
 - Gordon Huggins, TMC
 - Gavin Schnitzler, TMC
 - Joshua Ainsley, Tufts
 - Lionel Zupan, Tufts

Who am I?

Postdoc in Leon Reijmers' lab

Research focused on translation of mRNA in neuronal dendrites

Attended the CSHL Advanced Sequencing Technologies course in 2010

Practical experience in molecular biology and computational methodologies for RNA-Seq

Lecture Outline

- Acknowledgements
- Computational Biology Initiative
- **Course goals/outline**
- **Cluster test**
- Sequencing technologies
- Planning an RNA-seq experiment
- Library prep
- Design exercises
- Some helpful resources

Course Goals

You will learn how to:

- Prepare an RNA-Seq library
- Assess the quality of your RNA-Seq data
- Utilize basic Unix and R programming
- Align RNA-Seq reads to appropriate reference sequences
- Quantify gene and transcript expression
- Perform differential expression analysis

Course goals (for real)

You will have some idea
of where to start.

Course Outline

Day 1: Sequencing technologies, Library preparation, Experiment planning

Day 2: Unix introduction, File formats, Assessing sequence quality

Day 3: Sequence alignment, Read visualization, Expression quantification

Day 4: R introduction, Differential expression analysis

Day 5: Open workshop: Assembly, alternative splicing, RNA editing

Course Methodology

Lectures to introduce concepts

Hands on exercises – Participate!

Teamwork! Yes!

Minute cards

What did you learn?

What are you confused about?

Course website:

<http://sites.tufts.edu/cbi/resources/rna-seq-course/>

Why should you learn Unix/R?

Galaxy is a wonderful tool,
but it can't do everything...

Text manipulation

Adopt new methods

Biology is becoming a data-driven science

Bioinformatics is hard

Get used to being frustrated

Hundreds of poorly-maintained programs

Dozens of file formats

Inefficient tab delimited files

Papers without code

Programming is awesome

Access to newest technology and methods

Automation of tasks

Saves time, money, unnecessary collaborations

Don't like something? Fix it yourself!

Need something new? Make it yourself!

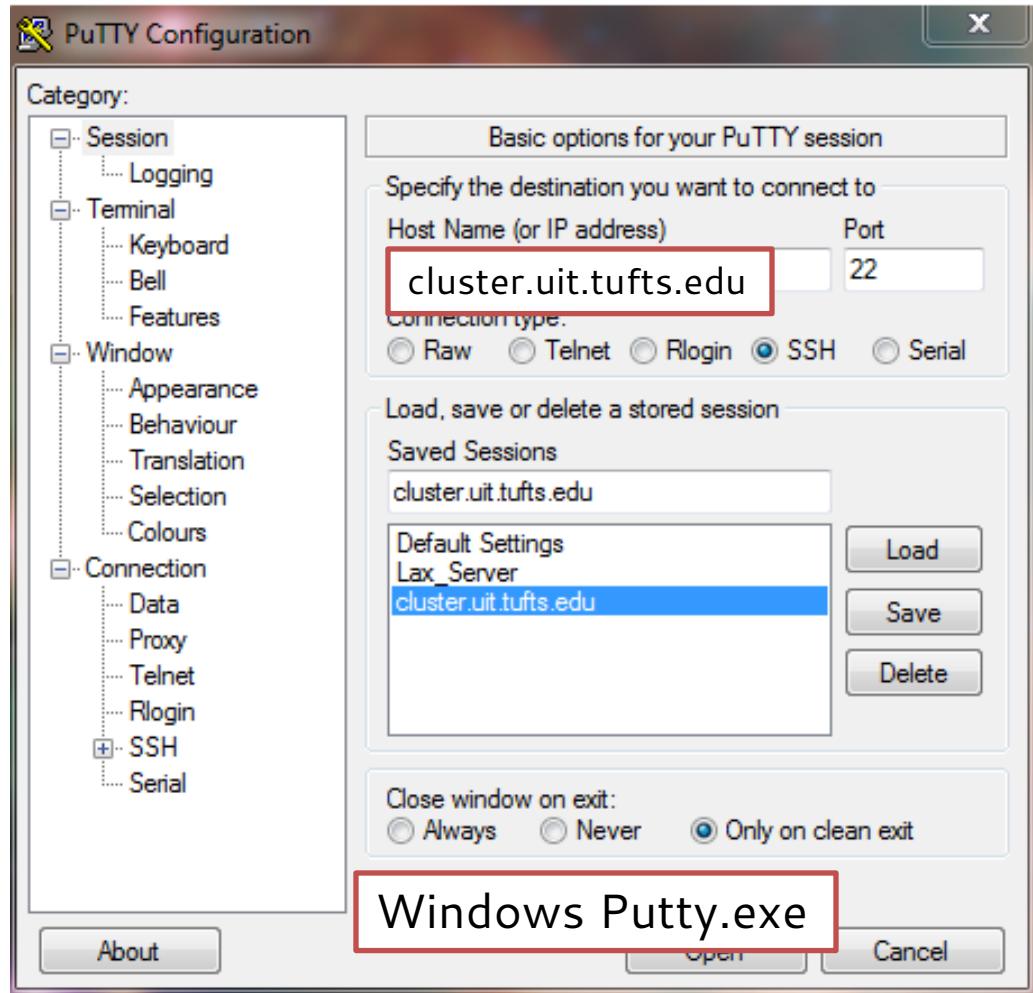
Amaze your friends and family (and interviewers!)

Bioinformatics job trends



Programming is an effective
tool you can use to ask the
scientific questions you are
interested in.

Login to Tufts Cluster



```
Brahma:~ lax$ ssh liyer01@cluster6.uit.tufts.edu
liyer01@cluster6.uit.tufts.edu's password:
Last login: Thu Jan 10 17:17:55 2013 from c-98-217-81-212.hsd1.ma.comcast.net
Red Hat Enterprise Linux Server release 6.2 (Santiago)
TAG: CLUSTER
GROUP: prod

CLUSTER.UIT.TUFTS
This is the user login node for Tufts University's NEW LINUX
Research Cluster running RedHat Enterprise Linux 6.
UNAUTHORIZED USE PROHIBITED
*****
2012 has marked further growth in demand for our High-Performance Computing
research cluster as well as the motivation to migrate from RHEL5 to RHEL6.
To facilitate this transition, CLUSTER6 was created to provide for the orderly
migration of cluster compute nodes to the most recent release of RedHat Linux.

*****
Questions? Send email to cluster-support@tufts.edu

By accessing and using this account and Tufts University computing resources,
you are responsible for adhering to the Tufts Responsible Use Policy at:
http://uit.tufts.edu/?pid=444 .

Please do not use /tmp for temporary or scratch file storage. All files of
this type should be kept in your directory in /scratch.

Users are required to use LSF to submit batch jobs to the compute nodes.

To run interactive GUI applications please see go to
http://go.tufts.edu/cluster and click on the "Application specific Information
FAQs" link.

To assess your user home directory disk usage, type: quota

For further information, please see http://go.tufts.edu/cluster
*****
[liyer01@tunic6 ~]$
```

ssh tuftsid@cluster.uit.tufts.edu

Integrated Development Environment for R

RStudio



The image shows the RStudio interface with four main components labeled:

- Source**: The top-left pane where R code is written.
- Console**: The bottom-left pane displaying the R startup message and command-line input.
- Workspace**: The top-right pane showing the R environment and datasets.
- Plots**: The bottom-right pane where plots and visualizations are displayed.

The RStudio interface includes a menu bar (File, Edit, Code, View, Plots, Session, Project, Build, Tools, Help), a toolbar, and a project browser.

```
R version 2.15.2 (2012-10-26) -- "Trick or Treat"
Copyright (C) 2012 The R Foundation for Statistical computing
ISBN 3-900051-07-0
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> |
```

Bioconductor.org

[Home](#)[Install](#)[Help](#)Search: [Developers](#)[About](#)[Home](#) » [Bioconductor 2.11](#) » [Software Packages](#) » [DESeq](#)

DESeq

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (2.11)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution

Author: Simon Anders, EMBL Heidelberg <sanders at fs.tum.de>

Maintainer: Simon Anders <sanders at fs.tum.de>

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("DESeq")
```

To cite this package in a publication, start R and enter:

```
citation("DESeq")
```

Documentation

[PDF](#) [R Script](#) Analysing RNA-Seq data with the "DESeq" package

[PDF](#) [R Script](#) vst.pdf

[PDF](#) Reference Manual

[Text](#) NEWS

Workflows »

Common Bioconductor workflows include:

- [Oligonucleotide Arrays](#)
- [High-throughput Sequencing](#)
- [Annotation](#)
- [Variants](#)
- [Flow Cytometry](#) and other assays
- [Finding Candidate Binding Sites for Known Transcription Factors via Sequence Matching](#)

Mailing Lists »

Post questions about Bioconductor packages to our mailing lists. Read the [posting guide](#) before posting!

- [bioconductor](#)
- [bioc-devel](#)

Lecture Outline

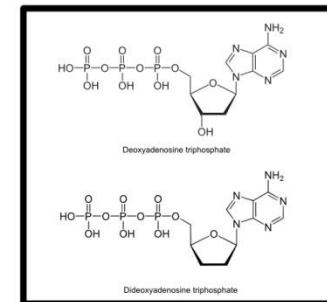
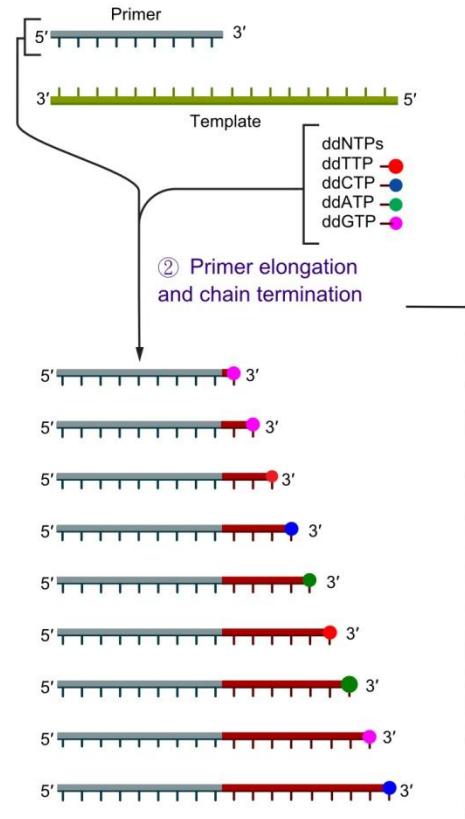
- Acknowledgements
- Computational Biology Initiative
- Course goals/outline
- Cluster test
- **Sequencing technologies**
- Planning an RNA-seq experiment
- Library prep
- Design exercises
- Some helpful resources

What is next generation sequencing?

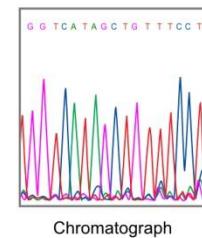
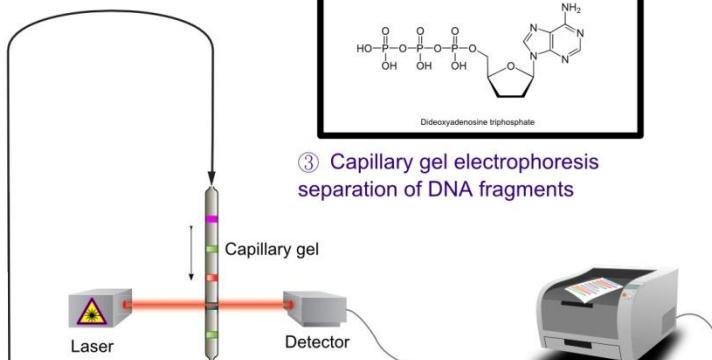
First generation = Sanger sequencing

① Reaction mixture

- Primer and DNA template
- ddNTPs with flourochromes
- DNA polymerase
- dNTPs (dATP, dCTP, dGTP, and dTTP)



③ Capillary gel electrophoresis separation of DNA fragments



④ Laser detection of flourochromes and computational sequence analysis

What is next generation sequencing?

Next generation = the stuff that came next

Next generation sequencing(NGS) =

High throughput sequencing(HTS) =

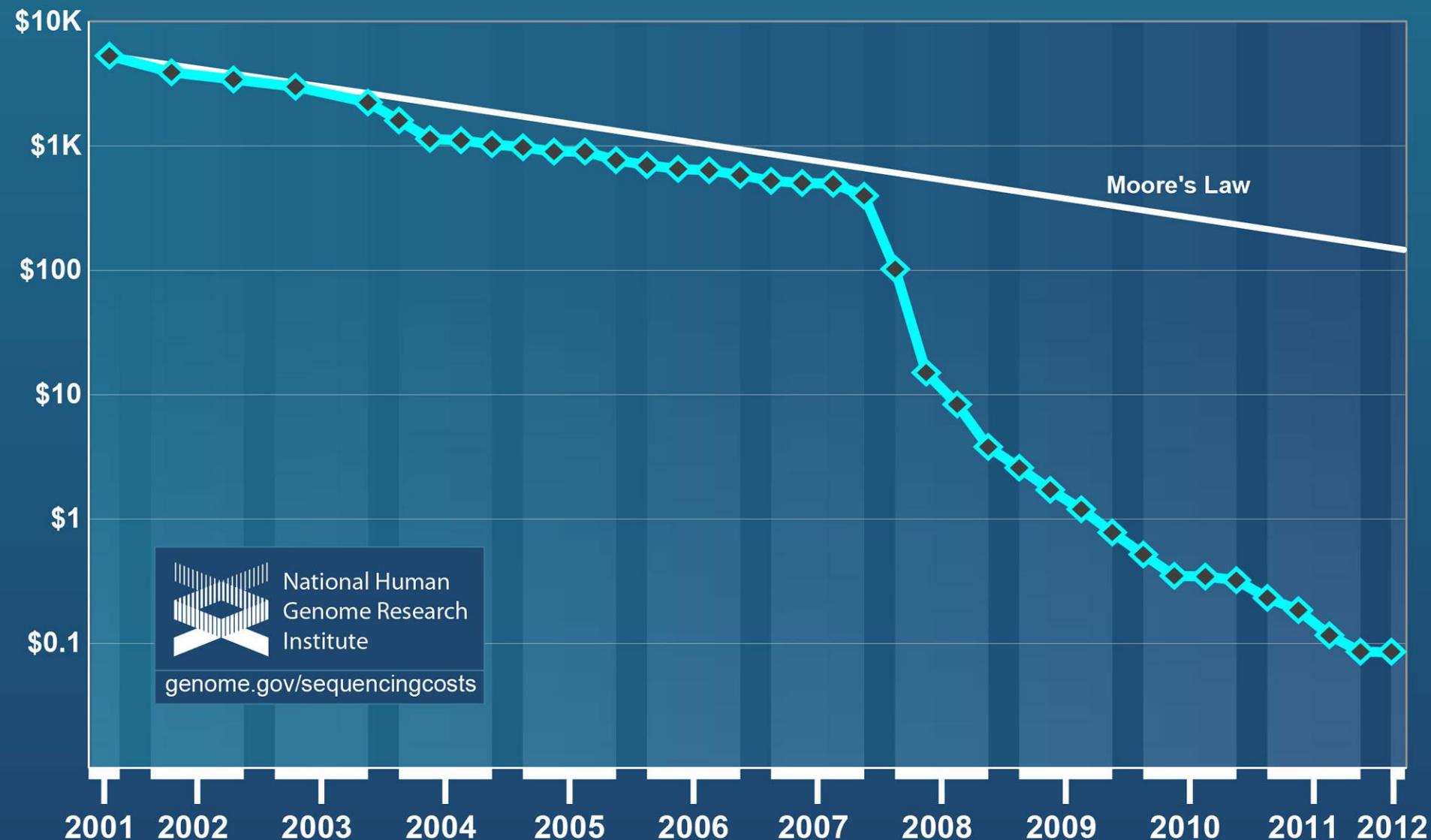
Deep sequencing

It's all just massively parallel sequencing.

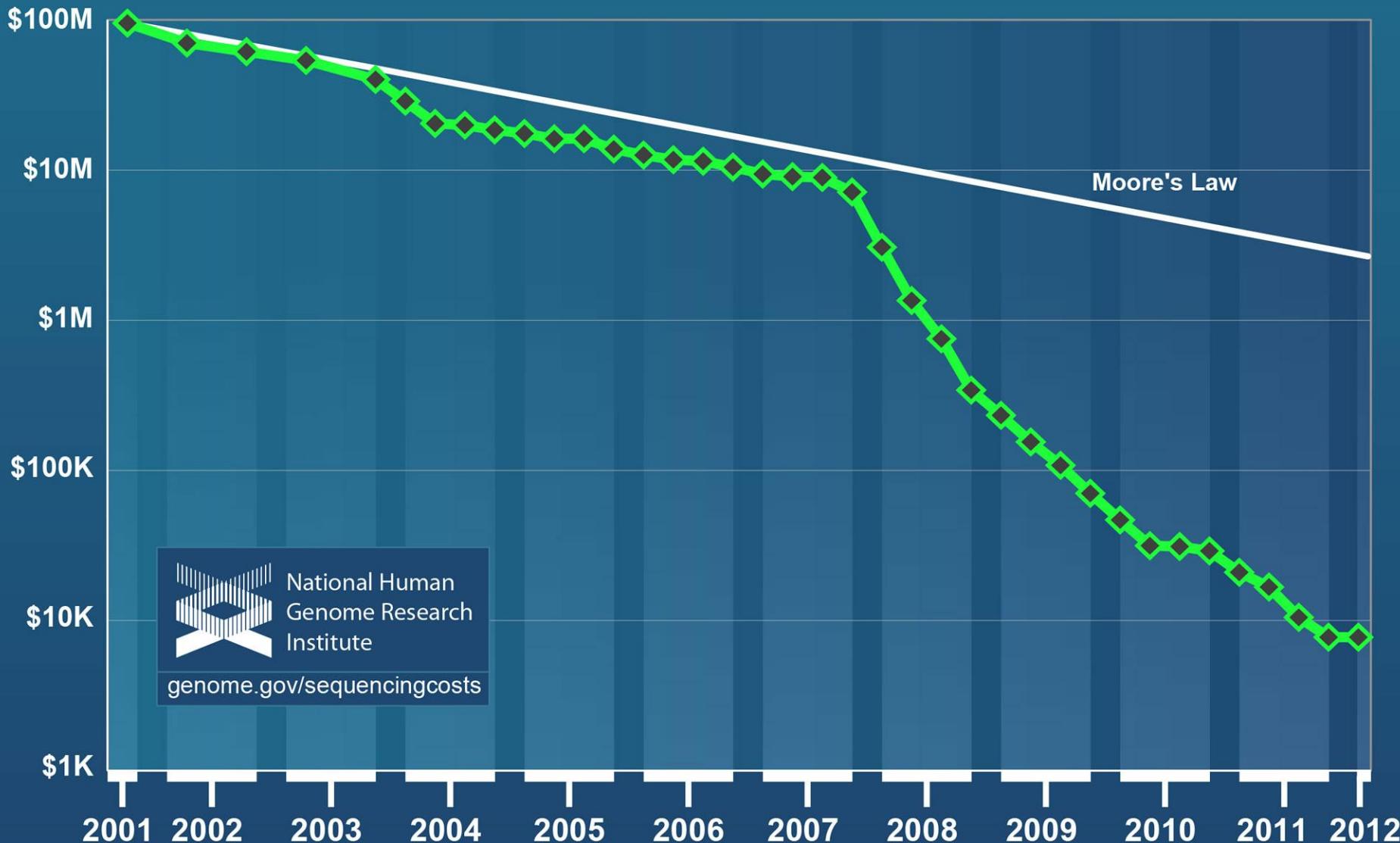
Second generation = Illumina/Solid/454

Third generation = Ion Torrent, Pacific Biosciences, Oxford Nanopore

Cost per Raw Megabase of DNA Sequence



Cost per Genome



National Human
Genome Research
Institute

genome.gov/sequencingcosts

HTS Commonalities (so far)

- Fragmentation of starting DNA
- Ligation with custom adapters
- Library amplification on a solid surface (either bead or glass)
- Direct detection of each incorporated nucleotide
- Hundreds of thousands to billions of reactions
- Shorter read lengths than capillary sequencers
- Count based data for quantitation
- Sampling both ends of every fragment sequenced (paired end reads)

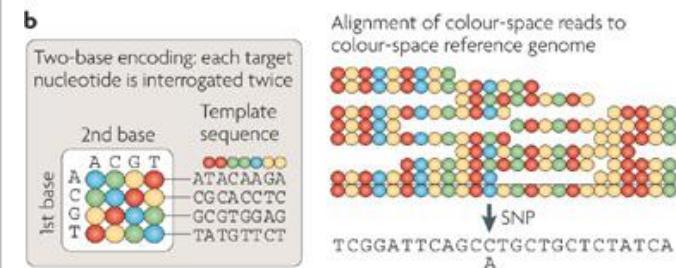
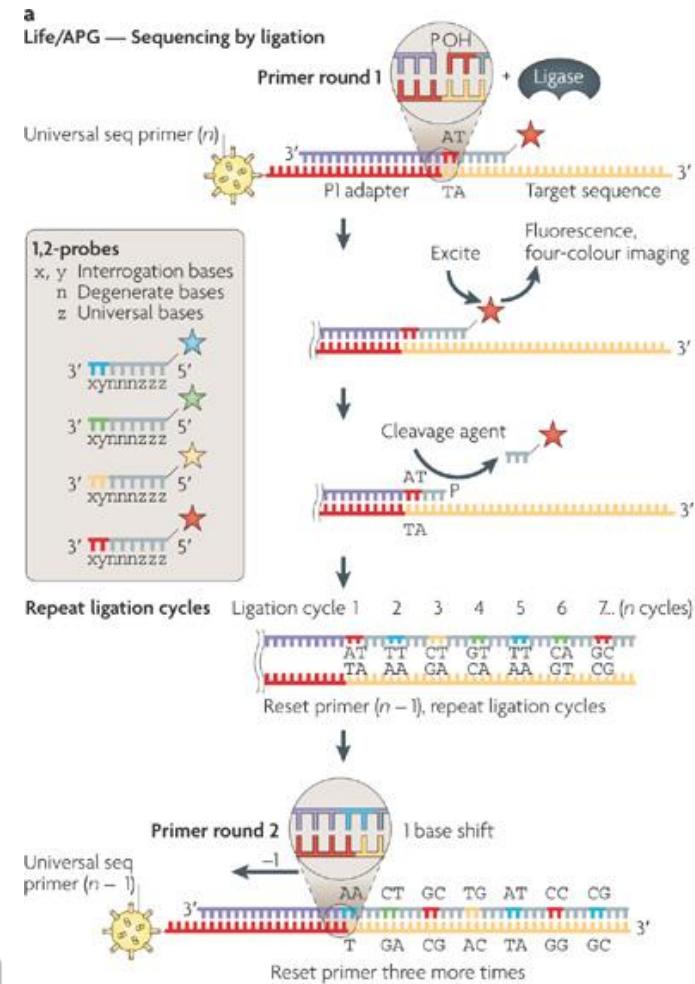
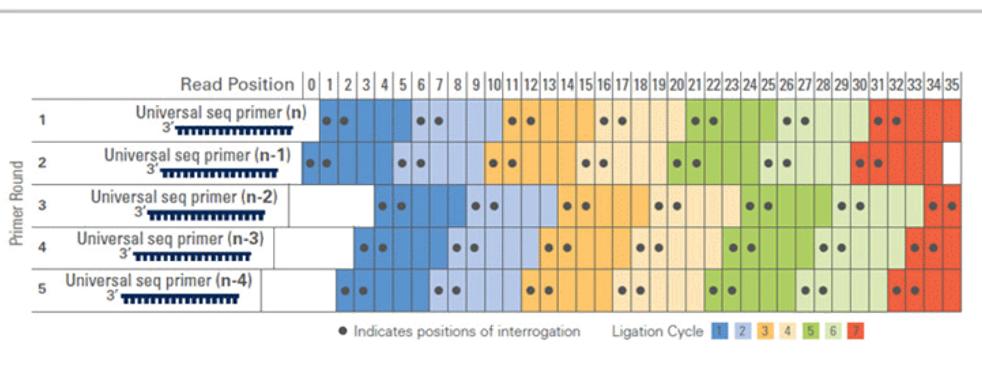
Sequencing Platforms

| Company | Platform | Amplification | Sequencing |
|-----------------|---------------------|---------------|--|
| Roche | 454 | emPCR | Synthesis, Pyrosequencing |
| Illumina | Illumina/ Solexa | Bridge PCR | Synthesis, Fluorescence |
| Life | SOLiD | emPCR | Ligation, Fluorescence |
| Life | Ion Torrent | emPCR | Synthesis, H ⁺ detection |
| PacBio | RS | None | Synthesis, ZMW fluorescence |
| Oxford Nanopore | ION | None? | Nanopore current flow |

SOLiD Sequencing

Sequencing-by-ligation (not so much anymore)

Shorter read length

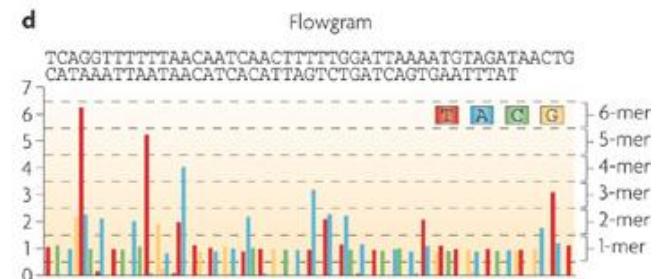
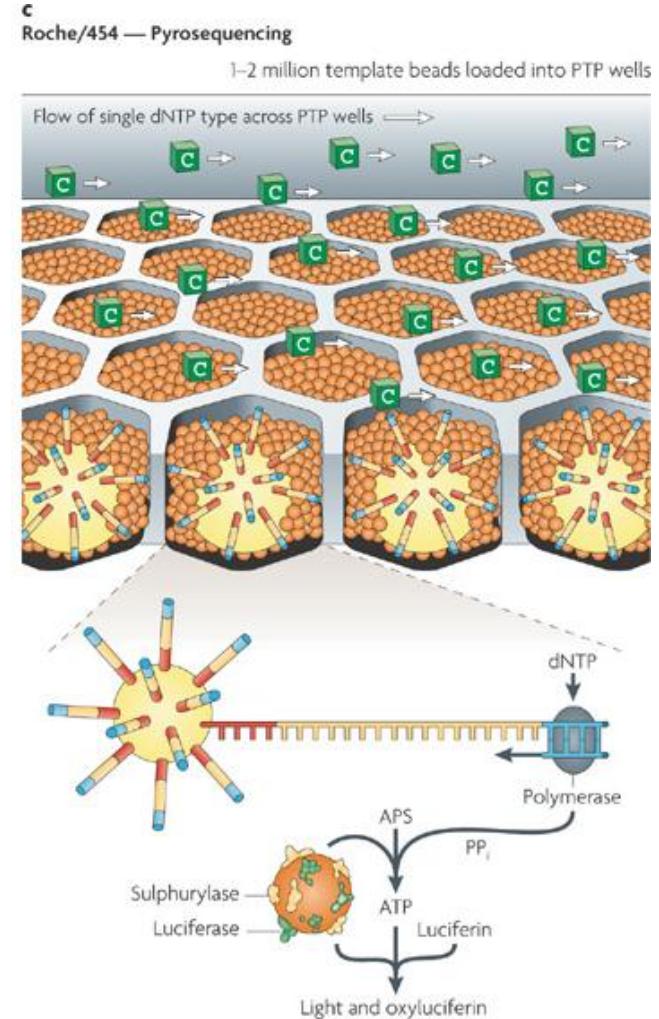


454 Sequencing

Pyrosequencing

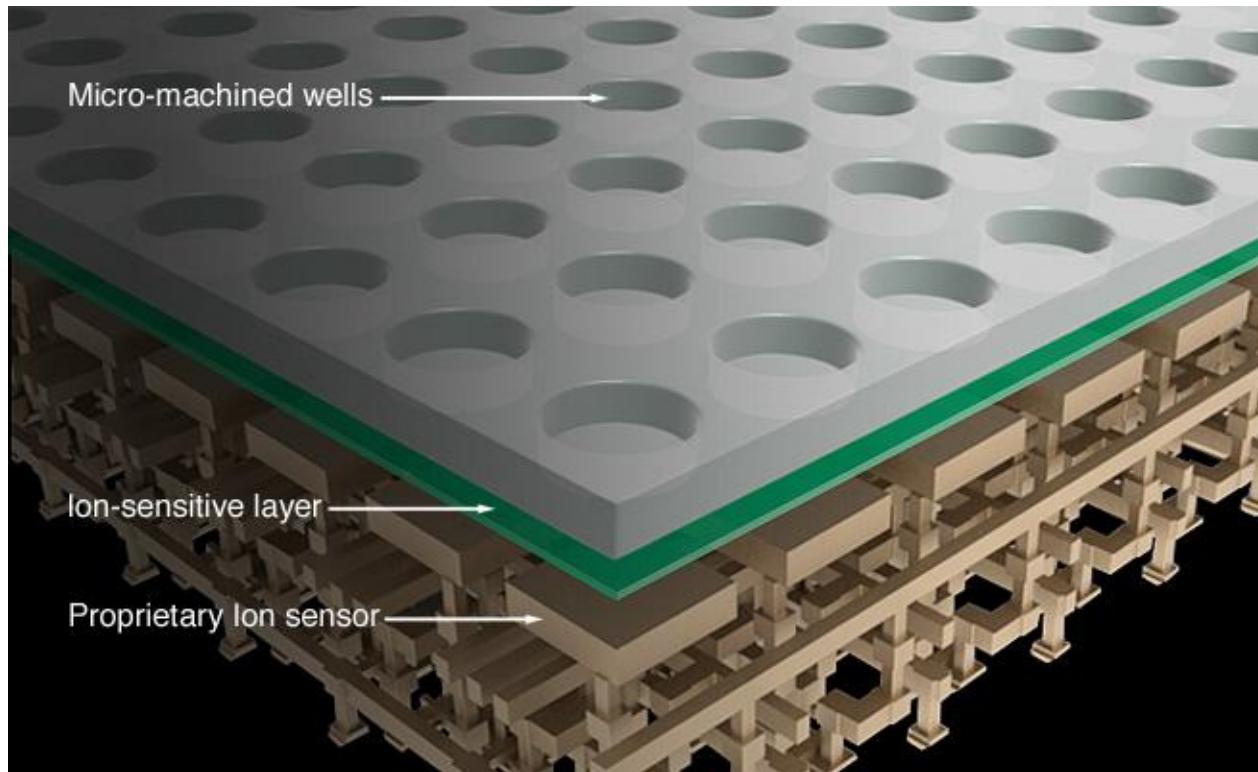
Read Length
(not so much
anymore)

Homopolymer errors



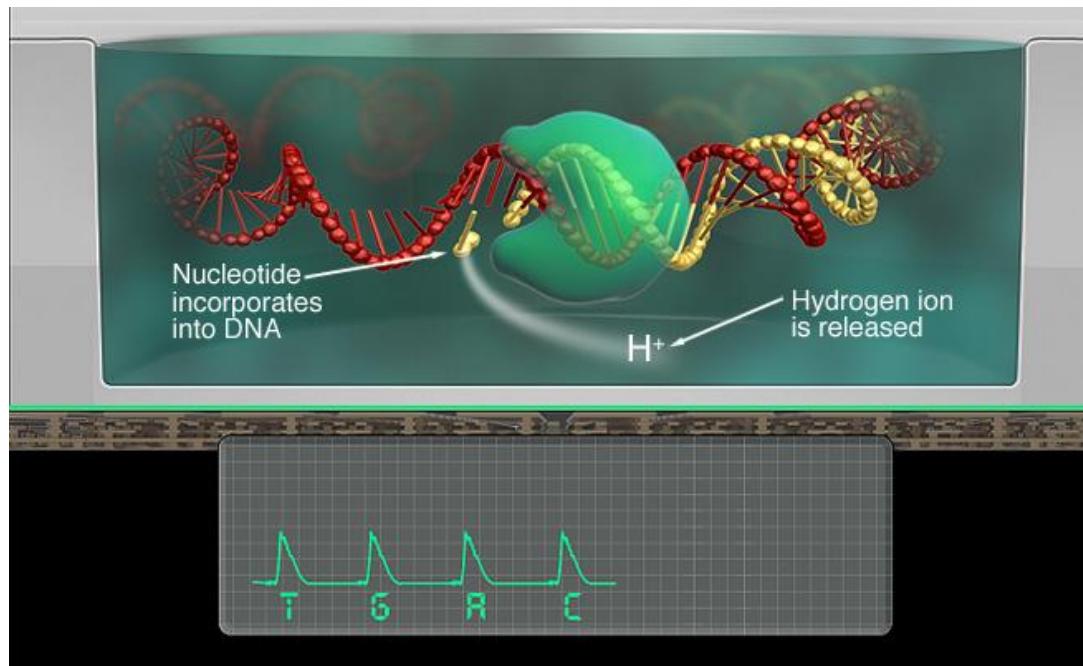
Ion Torrent

Sequencing by detecting the H⁺ ion released during nucleotide incorporation



Ion Torrent

Simple – unmodified nucleotides
added one at a time

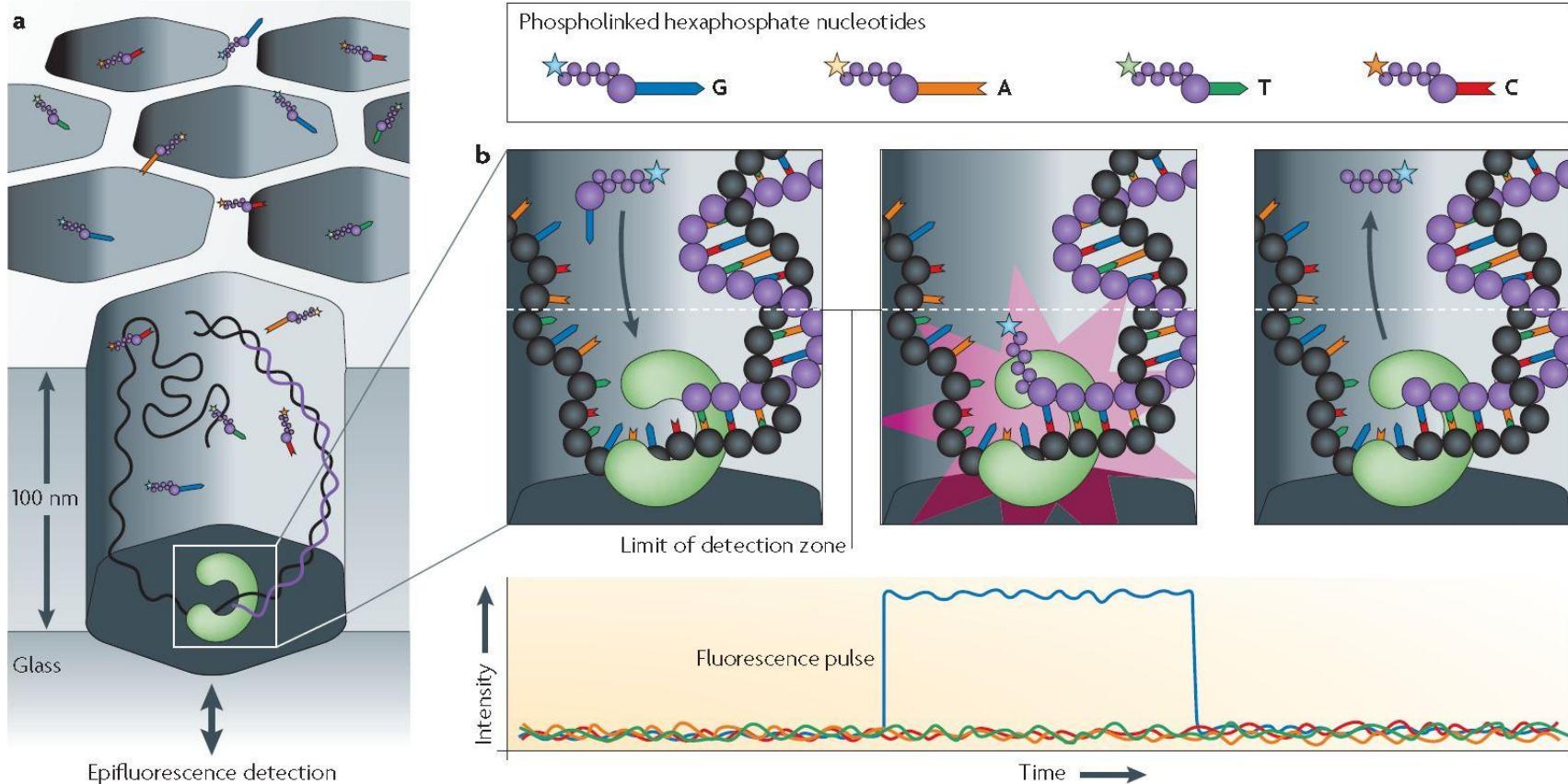


What error type do you think would be most common in Ion Torrent sequencing?

Pacific Biosciences

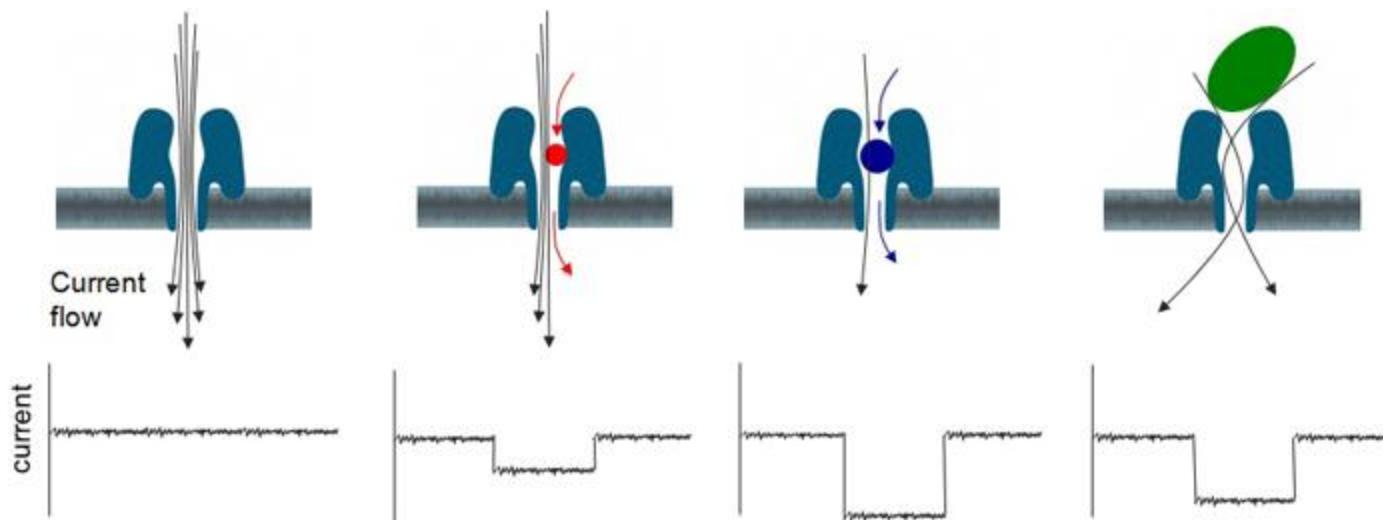
Real time, zero-mode waveguide sequencing
can detect modified bases

Pacific Biosciences — Real-time sequencing



Oxford Nanopore

Sequencing by detecting the change in ion flow that occurs when a nucleotide strand travels through a nanopore



[Video](#)

GridION and MinION



Illumina Sequencing

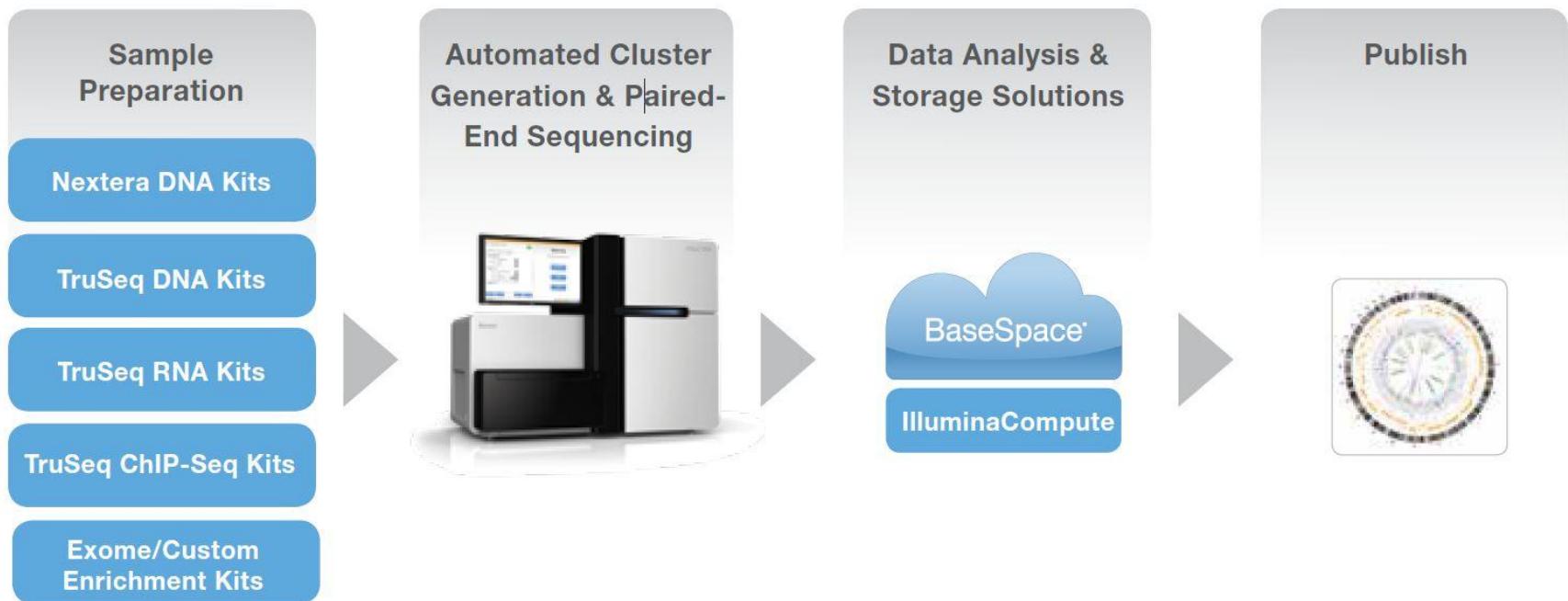
Most widely used sequencing platform

Highest throughput

Increasing accuracy, read length

Large number of third-party
and homebrew reagents

Illumina workflow



Illumina Platforms



HiSeq 2500/1500
Powerful. Flexible. Scalable.



HiSeq 2000/1000
Proven high-throughput platform.



HiScanSQ
Two proven technologies.
One powerful platform.



Genome Analyzer IIx
The most widely cited sequencing
platform.



MiSeq
The most accurate,
easiest-to-use benchtop
sequencer in the world.

| | | | | |
|--|---|---|---|--|
| Output (maximum) 600 Gb | 300 Gb | 150 Gb | 95 Gb | 7.8 - 8.5 Gb |
| Single Reads (maximum) 3 billion total 187 million/lane | 1.5 billion total 187 million/lane | 750 million total 94 million/lane | 320 million total 40 million/lane | 15 - 17 million total 15 - 17 million/lane |
| Paired-end Reads (maximum) 6 billion 374 million/lane | 3 billion 374 million/lane | 1.5 billion 188 million/lane | 640 million 80 million/lane | 30 - 34 million total 30 - 34 million/lane |
| Required input 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq |
| Read length 2 x 100 bp | 2 x 100 bp | 2 x 100 bp | 2 x 150 bp | 2 x 250 bp |
| Percentage of Bases > Q30 > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 100 bp) > 80% (2 x 150 bp) > 70% (2 x 250 bp) |

Illumina Platforms



- Clustering on-board
 - Fast Chemistry
 - Longer Reads
- Genome in a day
 - Clustering on-board
 - Complete walk-away workflow
 - Longer 2x150 reads
- Data rate
 - TDI scanning
 - Larger flow cell

Illumina 2500 run modes

1 Instrument – 2 Run Modes

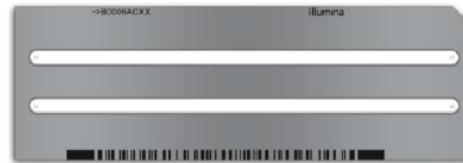
High Output Mode

600 Gb in ~10.5 days
3 billion clusters
cBot required



Rapid Run Mode

120Gb in ~1 day
600 million clusters
No cBot required



User configurable

6 human genomes
in 10.5 days

1 human genome
in a day



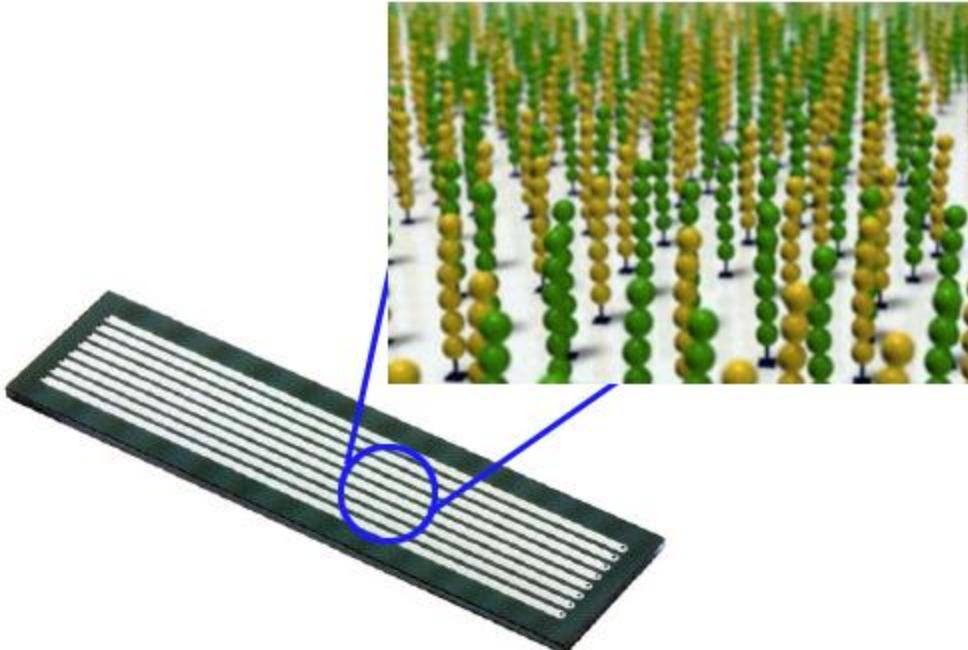
Highest output



Fastest turnaround

Illumina flowcells

microfluidic chambers
for reagent application



© CSD, Department of Biosystems Science and Engineering, ETH Zurich, Switzerland

Covalently linked oligos

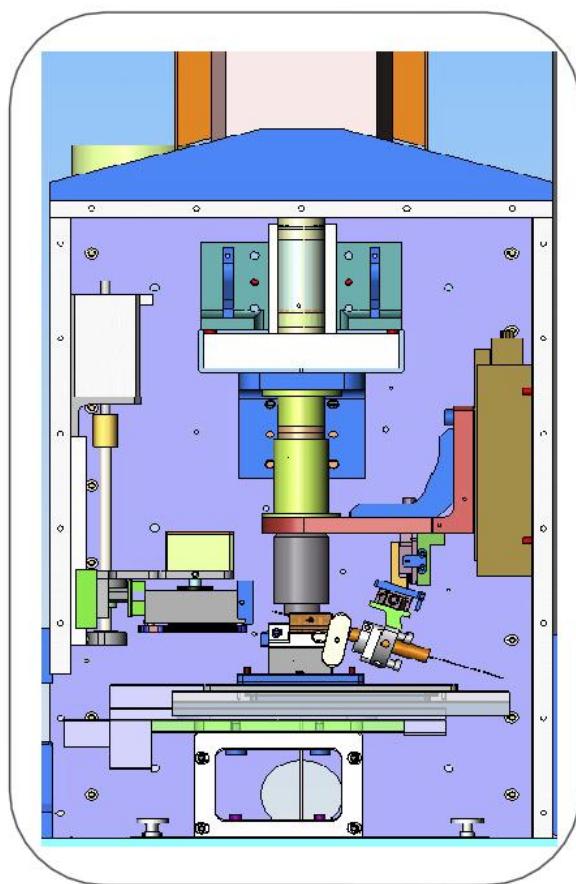
Invariant sample position

Throughput depends on
imaging area

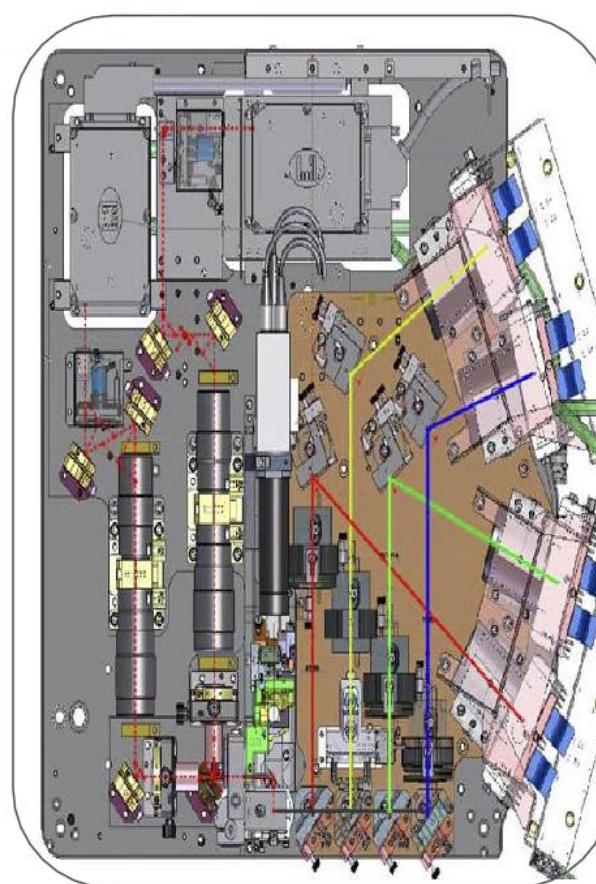
Increases in surface area
and imaging area

Illumina Technology

Fluorescent microscopy and microfluidics

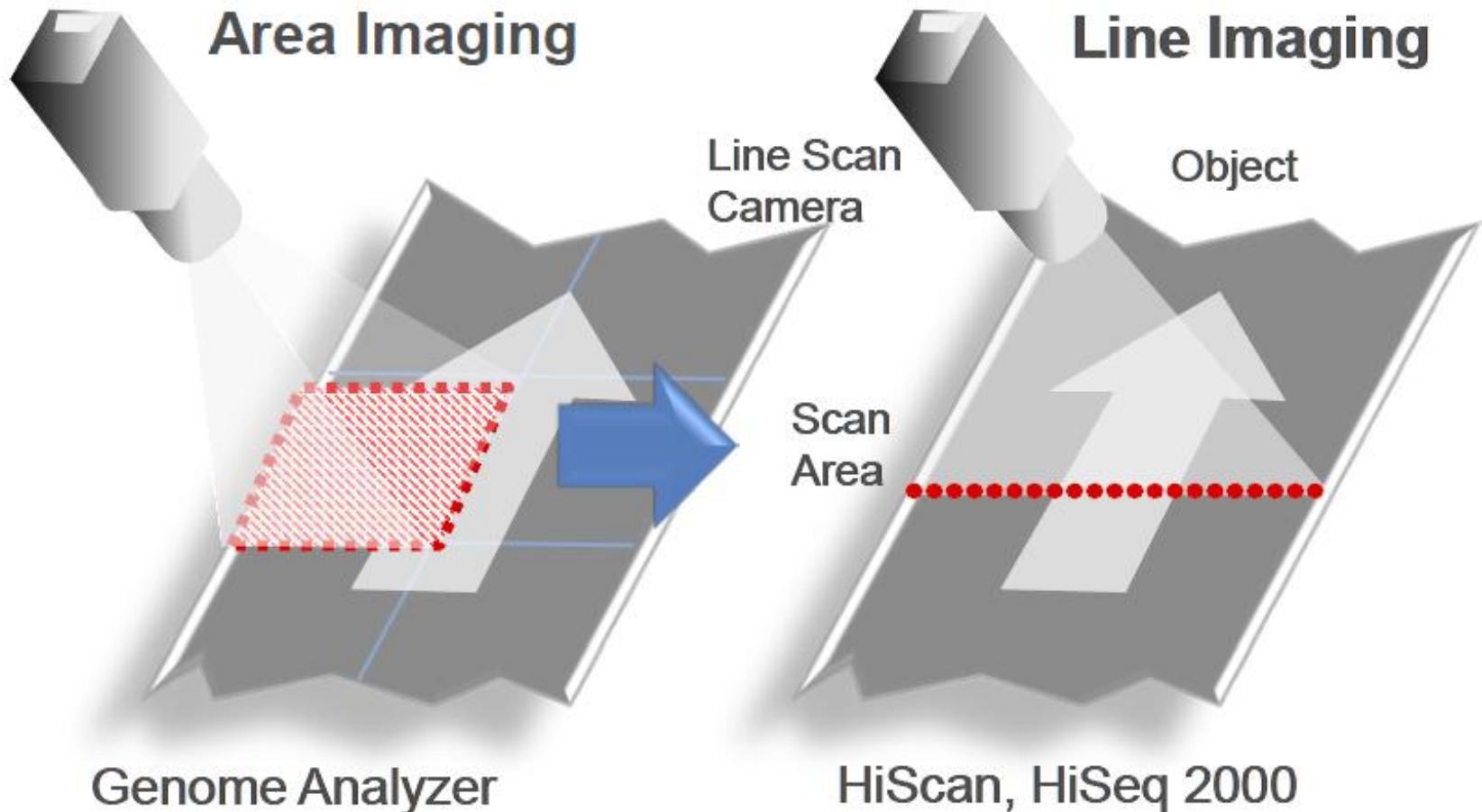


Genome Analyzer



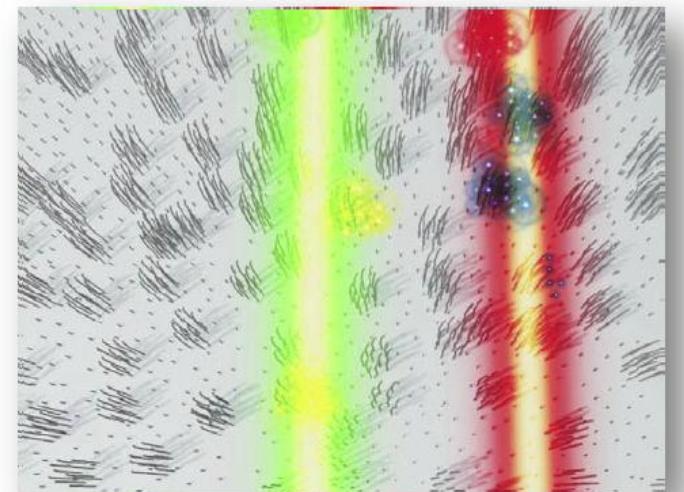
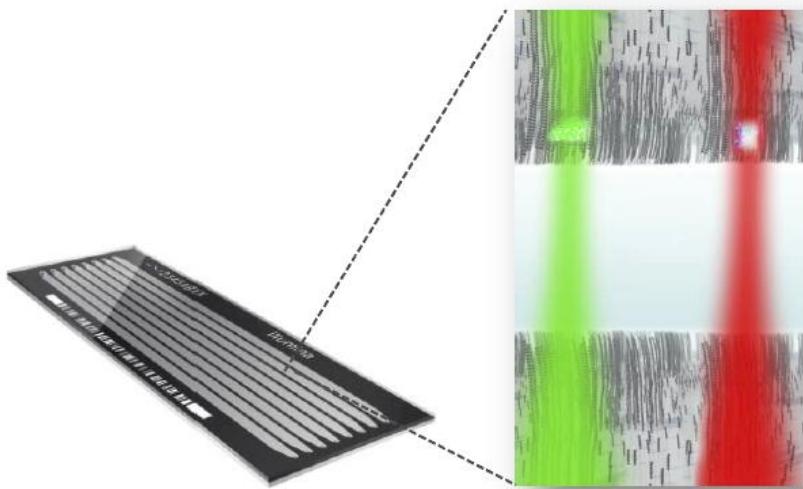
HiSeq 2000

Imaging an Illumina flowcell



Imaging an Illumina flowcell

HiSeq2000 – images bottom and top of flow cell sequentially



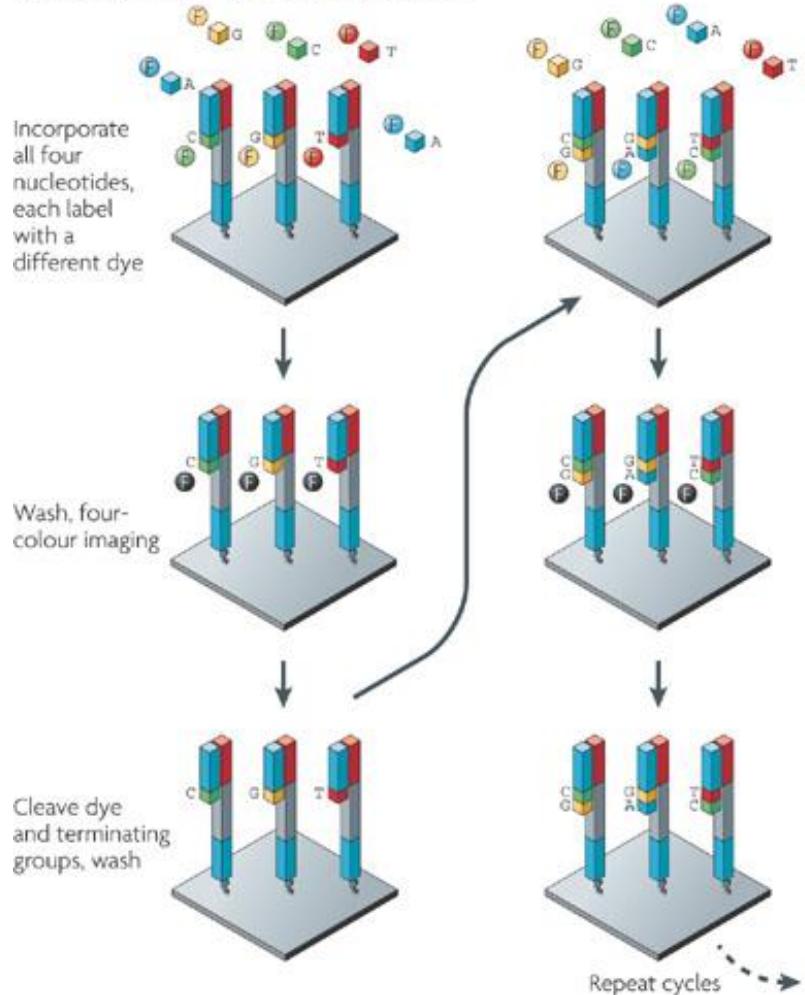
Illumina Sequencing

Sequencing-by-synthesis

Separate fluorescent tags on each nucleotide

Reversible terminators

a Illumina/Solexa — Reversible terminators



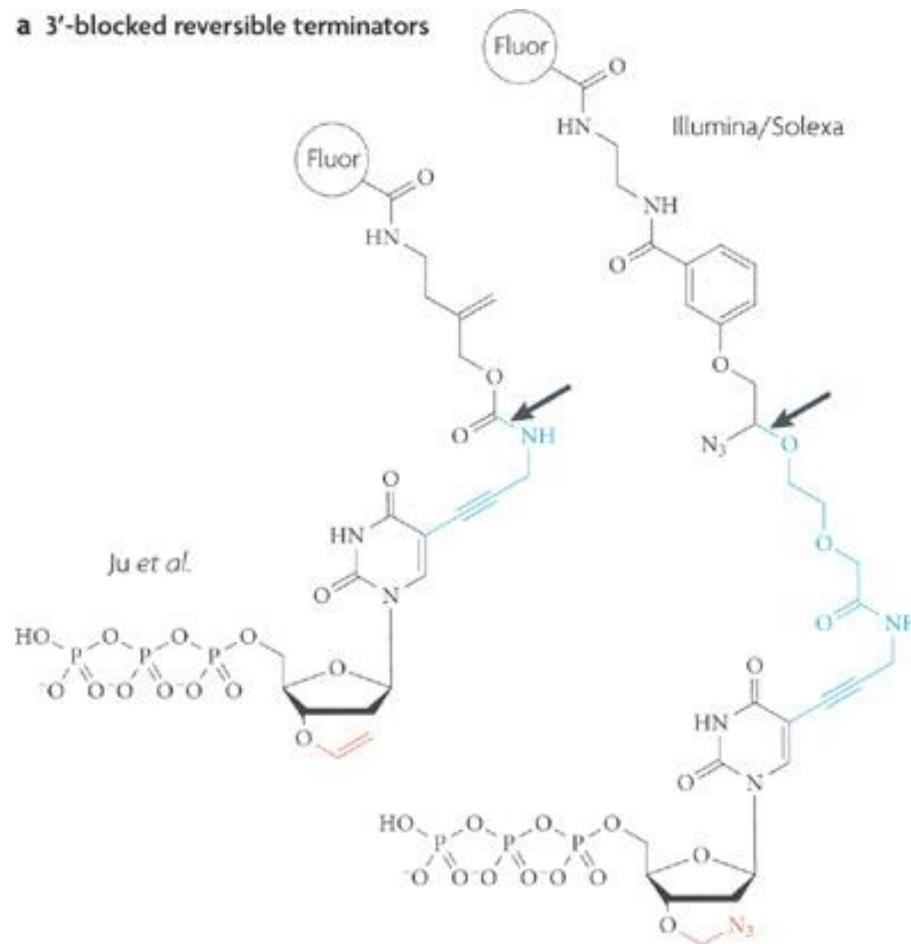
b



Top: CATCGT
Bottom: CCCCCC

Illumina chemistry – reversible terminators

a 3'-blocked reversible terminators



Library preparation steps

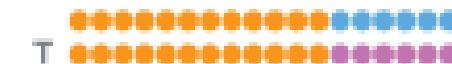
Fragmentation



End repair and A-tailing



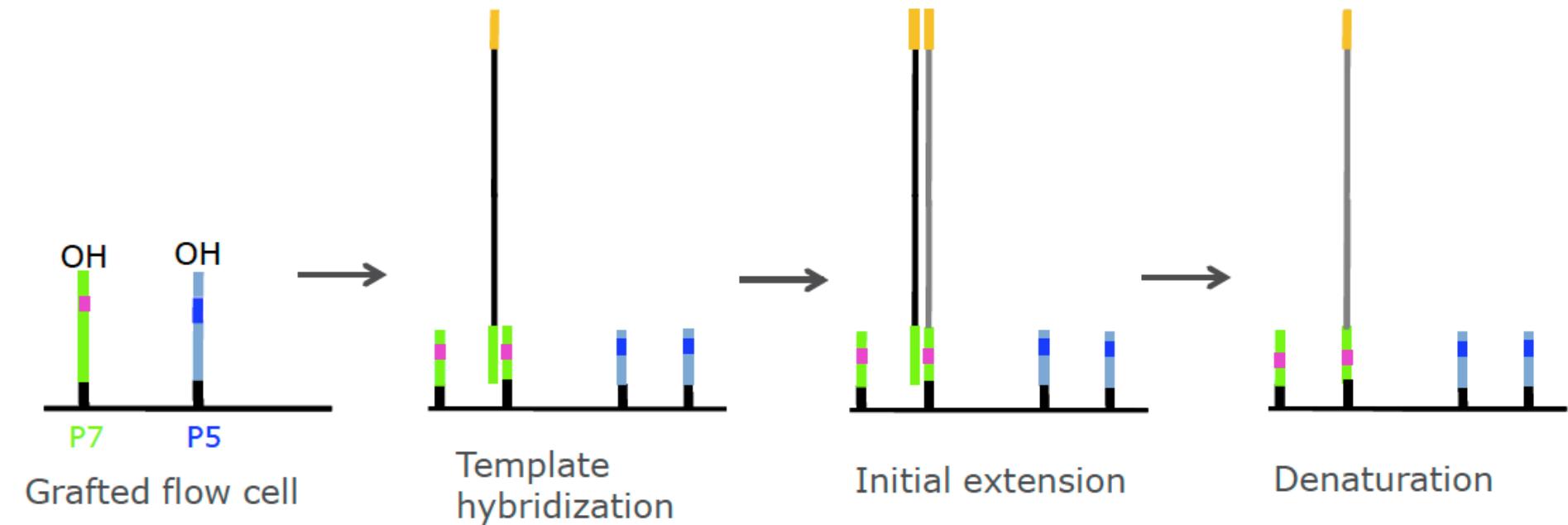
+



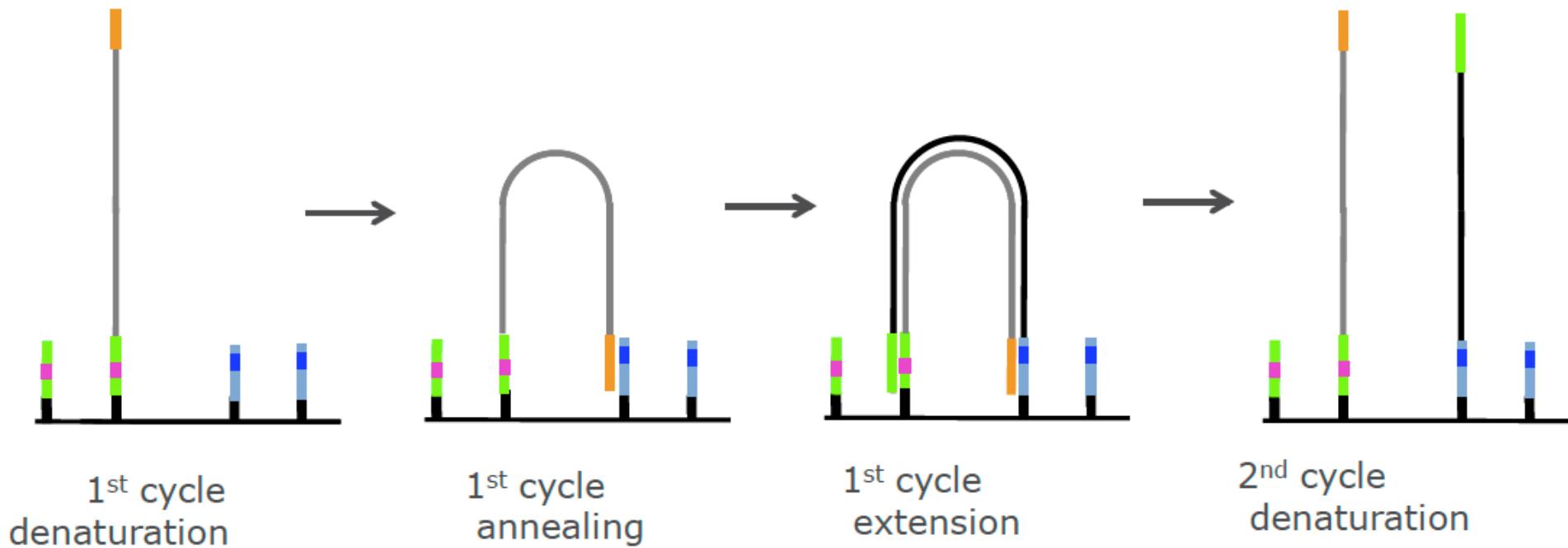
Adapter ligation



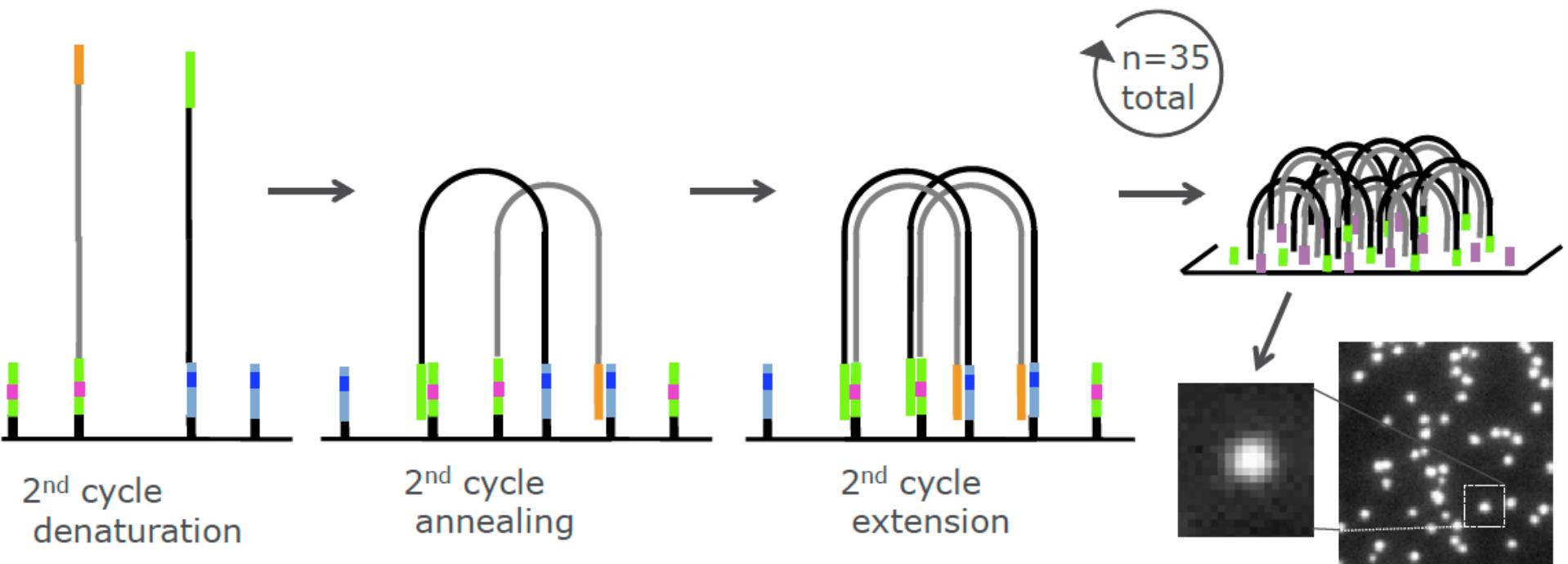
Getting your library on a flowcell



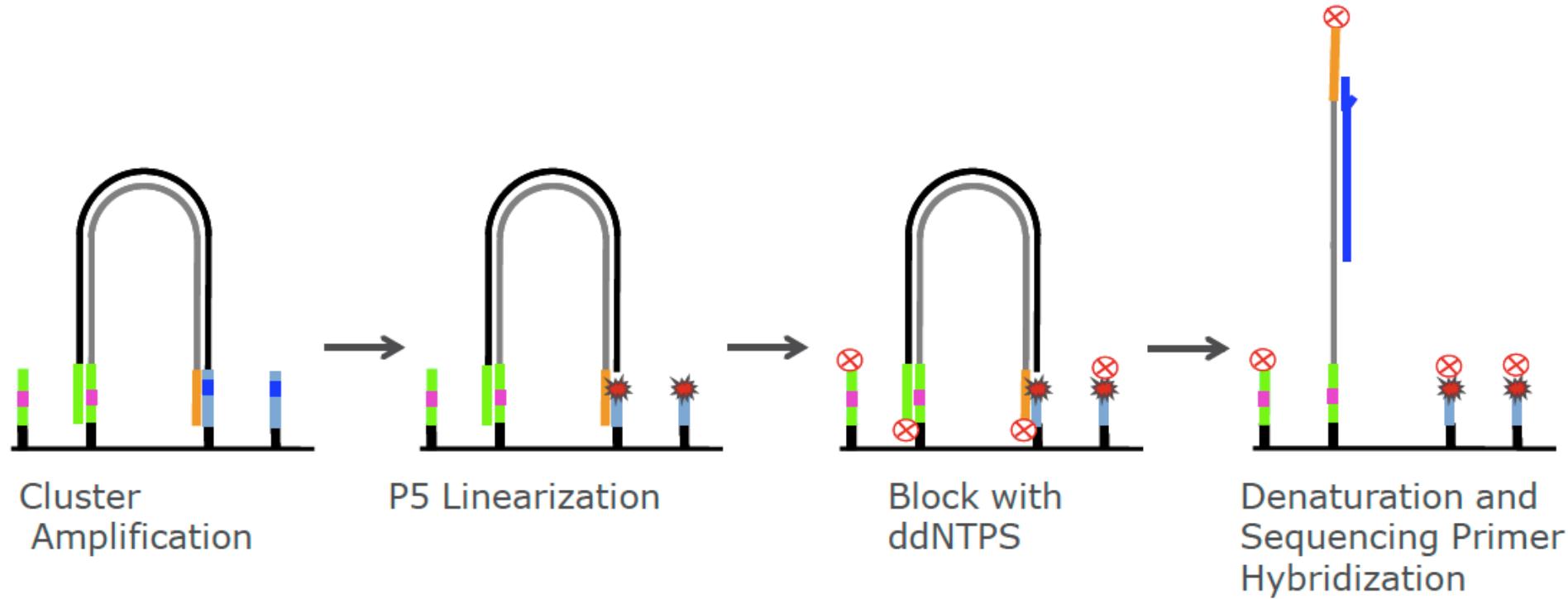
Cluster generation – Bridge amplification



Cluster generation – Bridge amplification



Sequencing



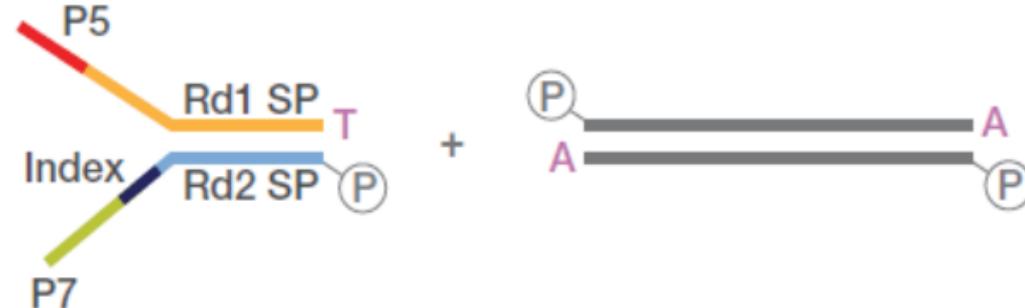
Single end – sequence one end
Paired end – sequence both ends

TruSeq Adapters

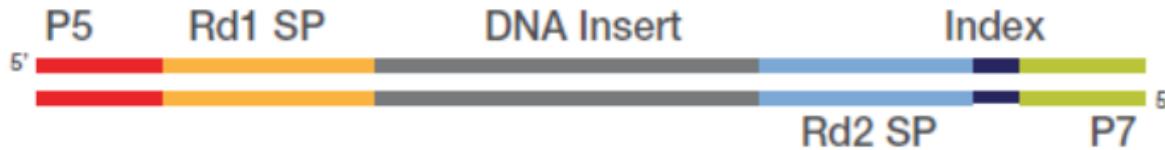
Forked adapters

Not complete
without PCR

Indexes/ barcodes
allow for multiplex
sequencing using a
third sequencing
read (currently up
to 24)

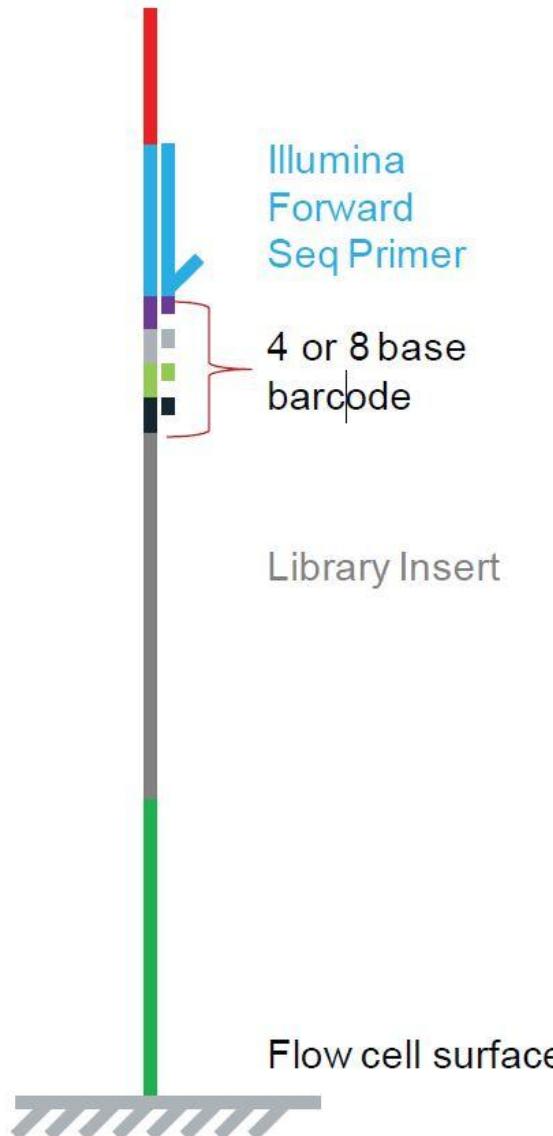


D. Ligate index adapter



E. Denature and amplify for final product

In-line multiplex adapters



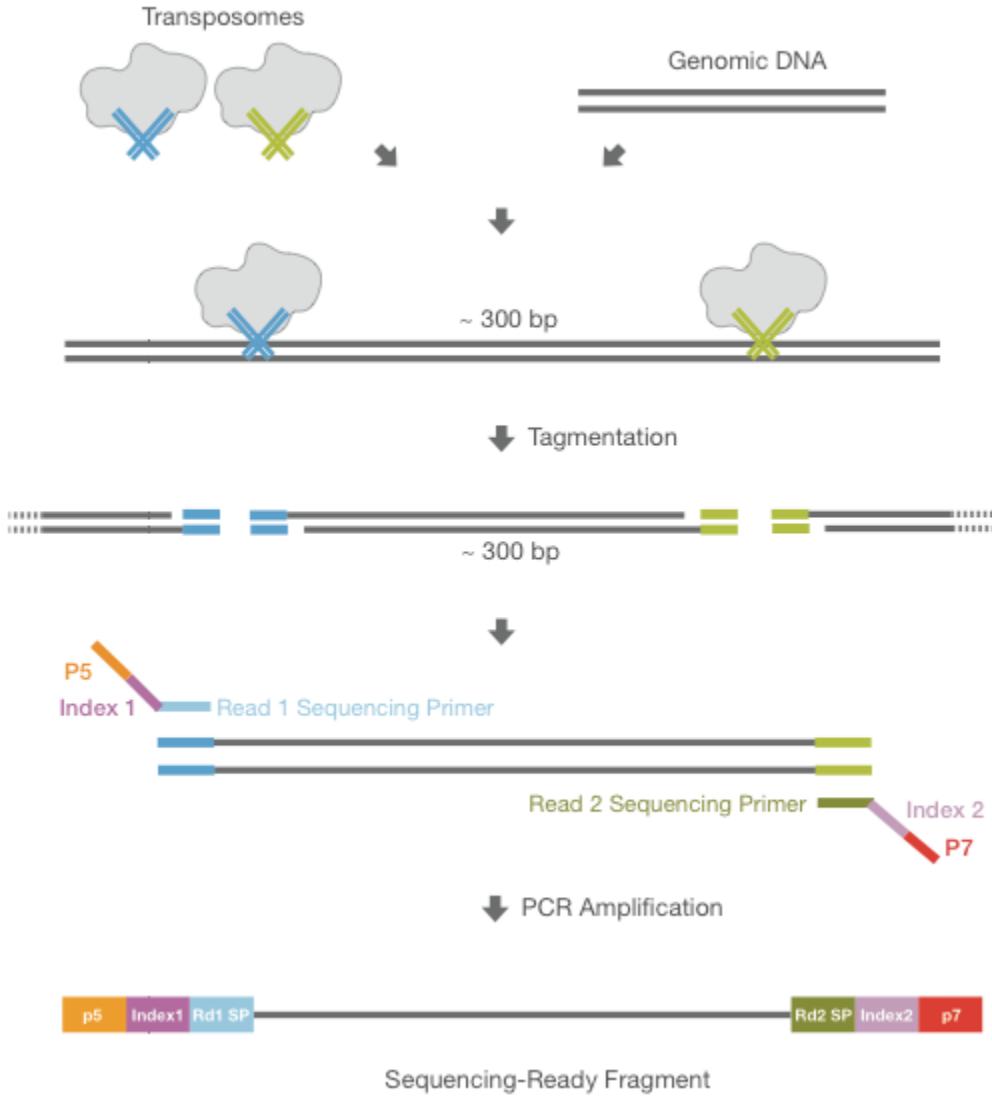
Indexes/barcodes allow for multiplex sequencing

Index/barcode is in the first 4–8 bases of the sequencing read

Allows for more indexes (currently up to 384)

Initial bases **MUST** be distributed randomly

NexTera “tagmentation”



Uses transposomes
to fragment DNA
and add adapters
simultaneously

Review

- Numerous technologies out there, with more in development.
- Illumina libraries need adapters to anneal properly to the flowcell.
- Illumina technology employs sequencing by synthesis, fluorescence imaging, and microfluidics.

Break

Minute cards!

Why are you taking this
course?

What do you want to learn?

What do you already know
about RNA-Seq?

Lecture Outline

- Acknowledgements
- Computational Biology Initiative
- Course goals/outline
- Cluster test
- Sequencing technologies
- **Planning an RNA-seq experiment**
- **Library prep**
- Design exercises
- Some helpful resources

ENCODE

(Encyclopedia of DNA elements)

“The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.”

30 papers published
simultaneously in September

ENCODE

Standards, Guidelines and Best Practices for RNA-Seq

Detailed document describing desirable characteristics of an RNA-Seq experiment

Not comprehensive (designed for cell lines)

Not necessary to follow all the guidelines,
depending on your experiment

Planning an RNA–Seq experiment

- **How will I make the cDNA?**
- How will I remove rRNA, gDNA, and mitochondrial sequences?
- How will I prepare the library?
- How many PCR cycles?
- How much sequencing should I do?
- Should I align or assemble?
- What should I align to?
- How do I quantify expression?
- How do I test for differential expression?

It depends...

Source RNA

Two main concerns:

Amount (best assessed with Ribogreen)

≥100 ng, use any kit

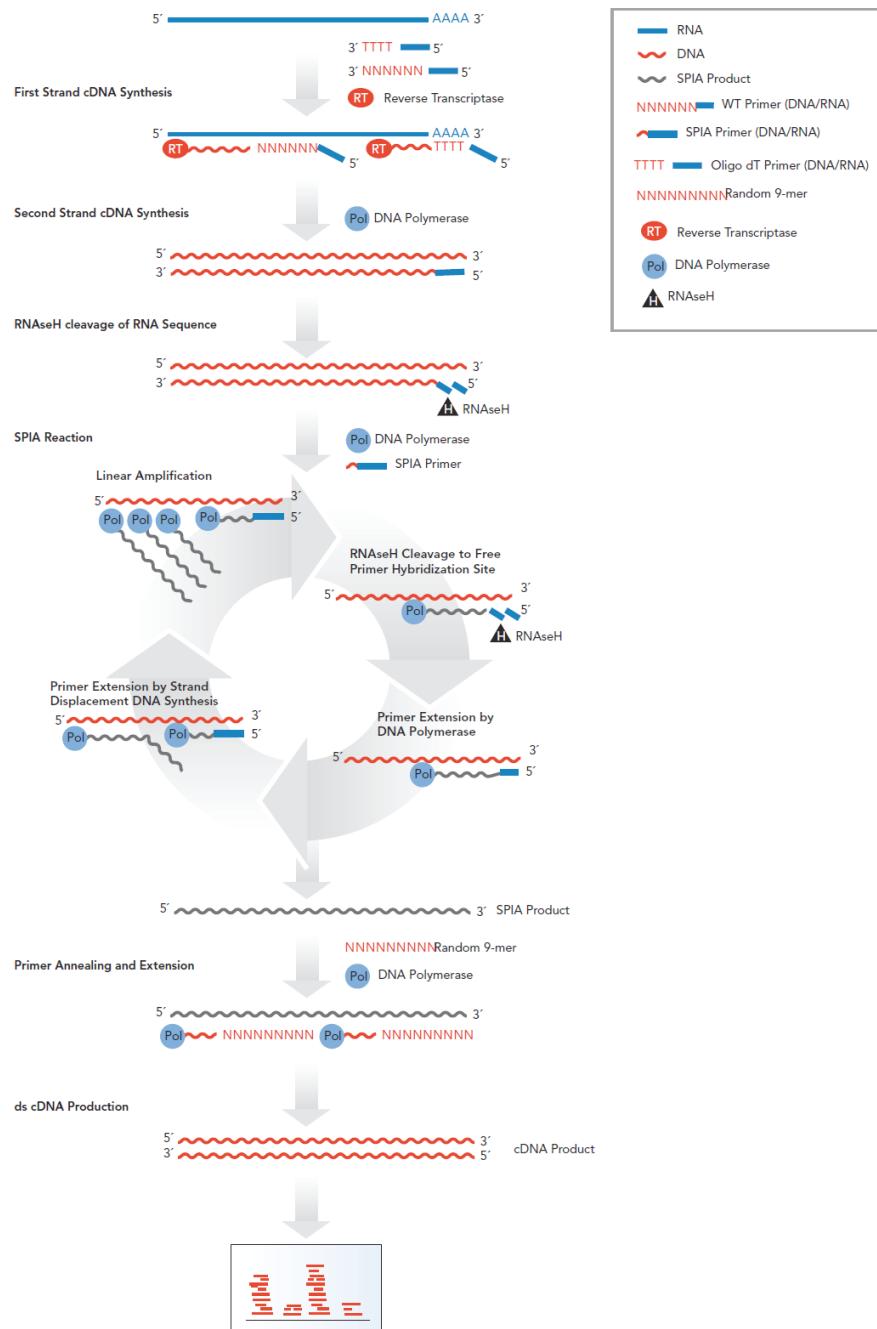
<100 ng, could require amplification

Quality (best assessed with Bioanalyzer)

High RIN, use any kit

Low RIN, RT with random hexamers

Schematic Ovation® RNA-Seq Process

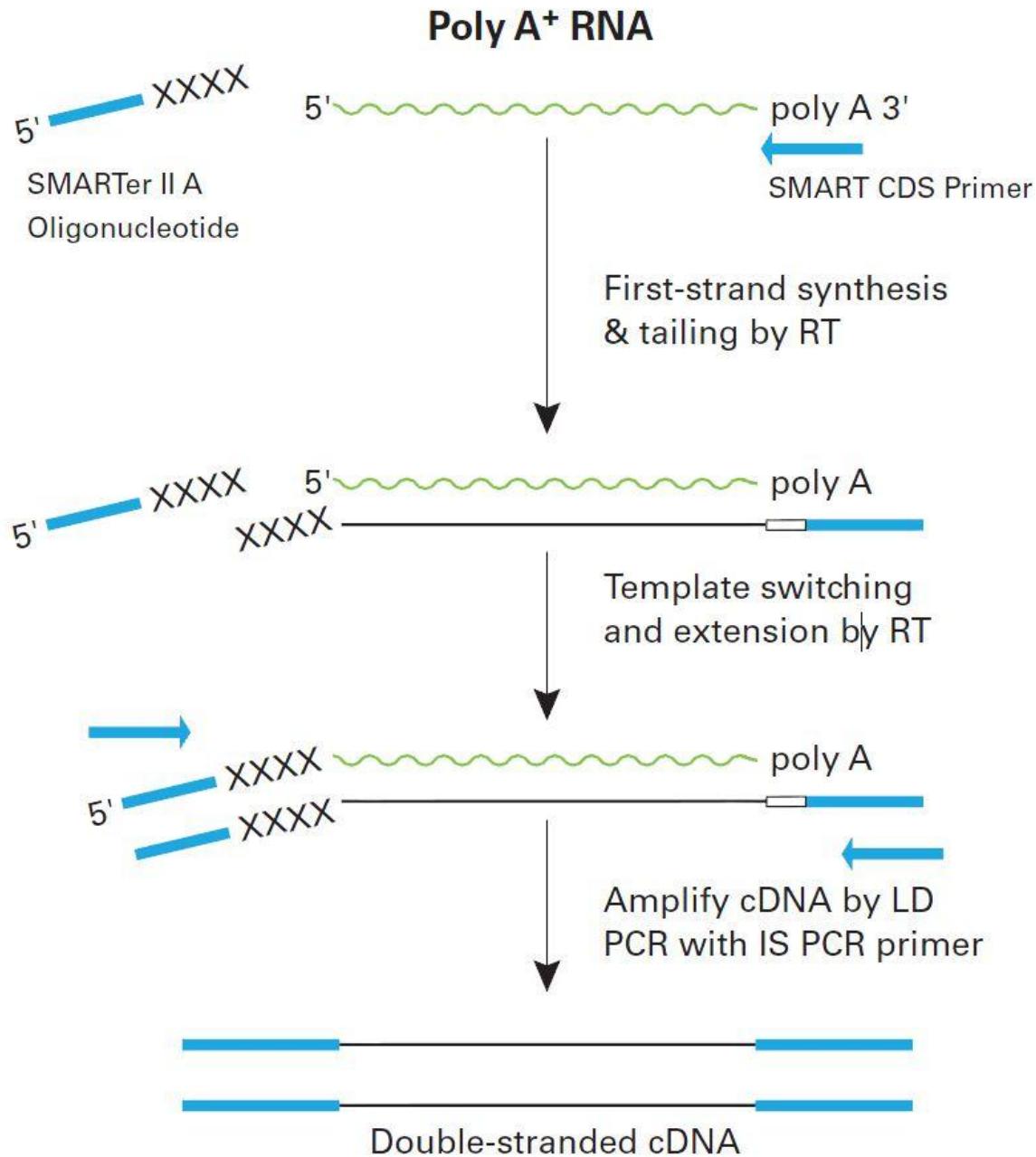


Nugen Ovation

Linear amplification

Works with
fragmented
samples

Allows sub-
nanogram input
(hundreds of pgs)



Clontech
SMARTer
Ultra Low

Requires high quality RNA

Uses oligo dT priming for RT

Allows single cell input (10pg)

Reverse Transcription

Two main options for reverse priming:

Oligo dT

requires high quality RNA

can produce an AT bias and a 3' bias

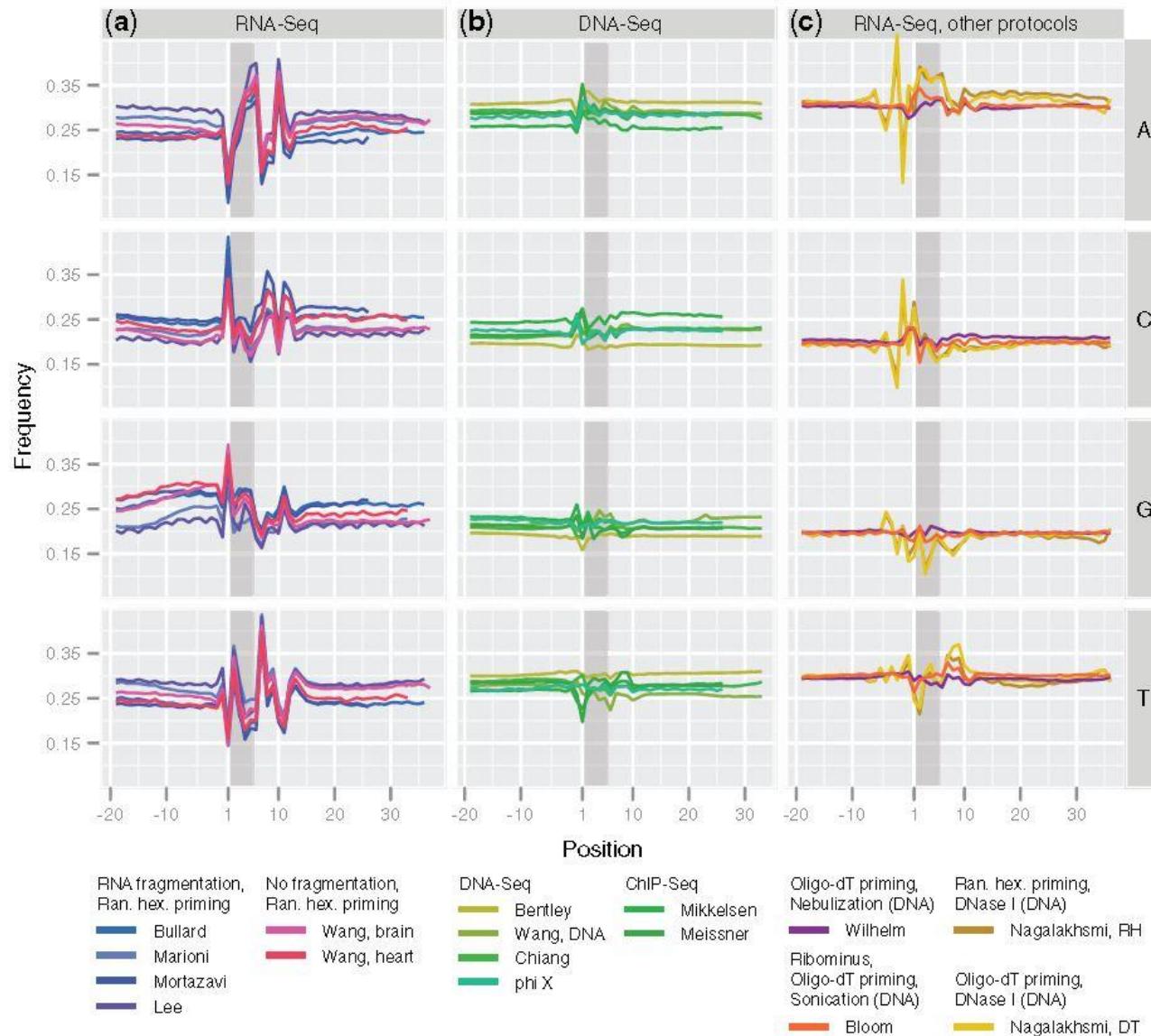
Random primers

works with fragmented RNA

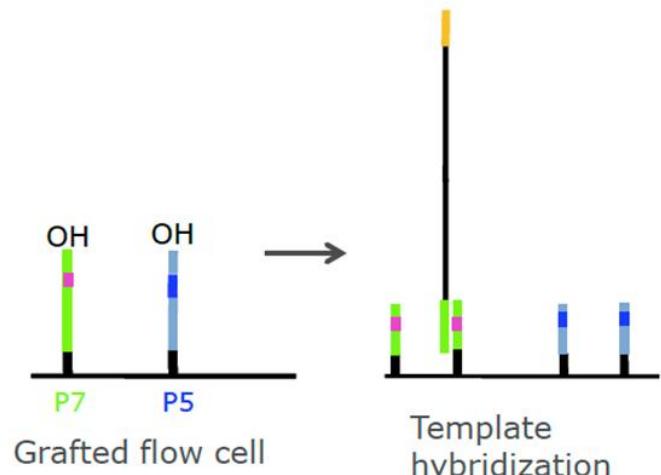
not really random; produces a bias at the beginning of the read

Use a mixture of both

Random Hexamer bias

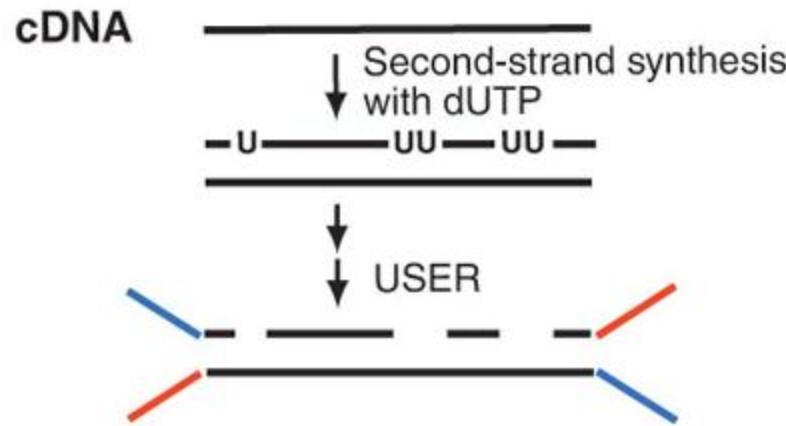


Strand specific sequencing



Standard RNA-Seq does not give strand information

Need a separate step to distinguish between strands



Useful for more complicated genomes with overlapping genes or for discovery

Levin, J.Z., et.al.(2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Meth

Planning an RNA–Seq experiment

- How will I make the cDNA?
- **How will I remove rRNA, gDNA, and mitochondrial sequences?**
- How will I prepare the library?
- How many PCR cycles?
- How much sequencing should I do?
- Should I align or assemble?
- What should I align to?
- How do I quantify expression?
- How do I test for differential expression?

Removing genomic DNA and mtDNA

DNase treatment if at all possible

RNA-Seq is a **sampling experiment!**

The more unwanted sequences you have in your sample, the fewer useful reads you will get in the end.

Removing ribosomal RNA

rRNA is >90% of the RNA in a cell!

polyA selection

Collect polyadenylated RNAs

oligo dT beads

requires high quality RNA

can introduce 3' bias, AT bias

no organisms without polyadenylation

Removing ribosomal RNA

Subtractive hybridization

Pull out rRNA with specific probes

Preserves non-polyadenylated RNA

Products not available for all species

Best choice for bacteria

Removing ribosomal RNA

Not so random primers

Primers for RT are screened against rRNA sequences

Works for polyA and non-polyA

Increased sequence bias

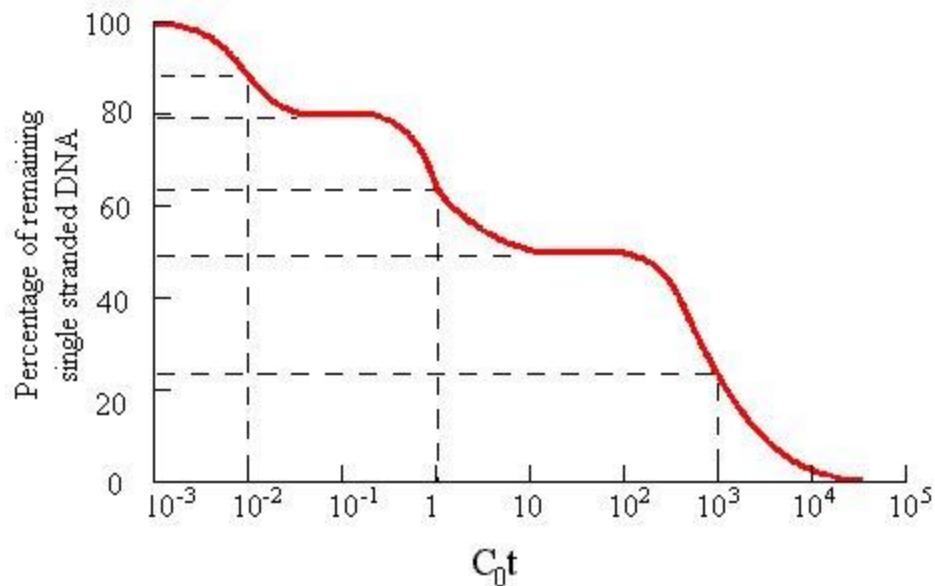
Removing ribosomal RNA

DSN (double strand nuclease) normalization

Very tricky to
optimize

Not for DE

Good choice mainly
for organisms with
no reference



Removing ribosomal RNA

Selective hybridization/PCR amplicons

Select known target genes

You pick what you want to sequence!

Not for discovery

Thoroughly developed hypothesis

Can be expensive

Need to be careful about low complexity

Planning an RNA–Seq experiment

- How will I make the cDNA?
- How will I remove rRNA, gDNA, and mitochondrial sequences?
- **How will I prepare the library?**
- How many PCR cycles?
- How much sequencing should I do?
- Should I align or assemble?
- What should I align to?
- How do I quantify expression?
- How do I test for differential expression?

RNA or cDNA fragmentation

Fragmentation must occur at some point

The larger the insert, the
lower the clustering efficiency

100–300 bp insert size is best

Fragment size = size of insert + adapters (120bp)

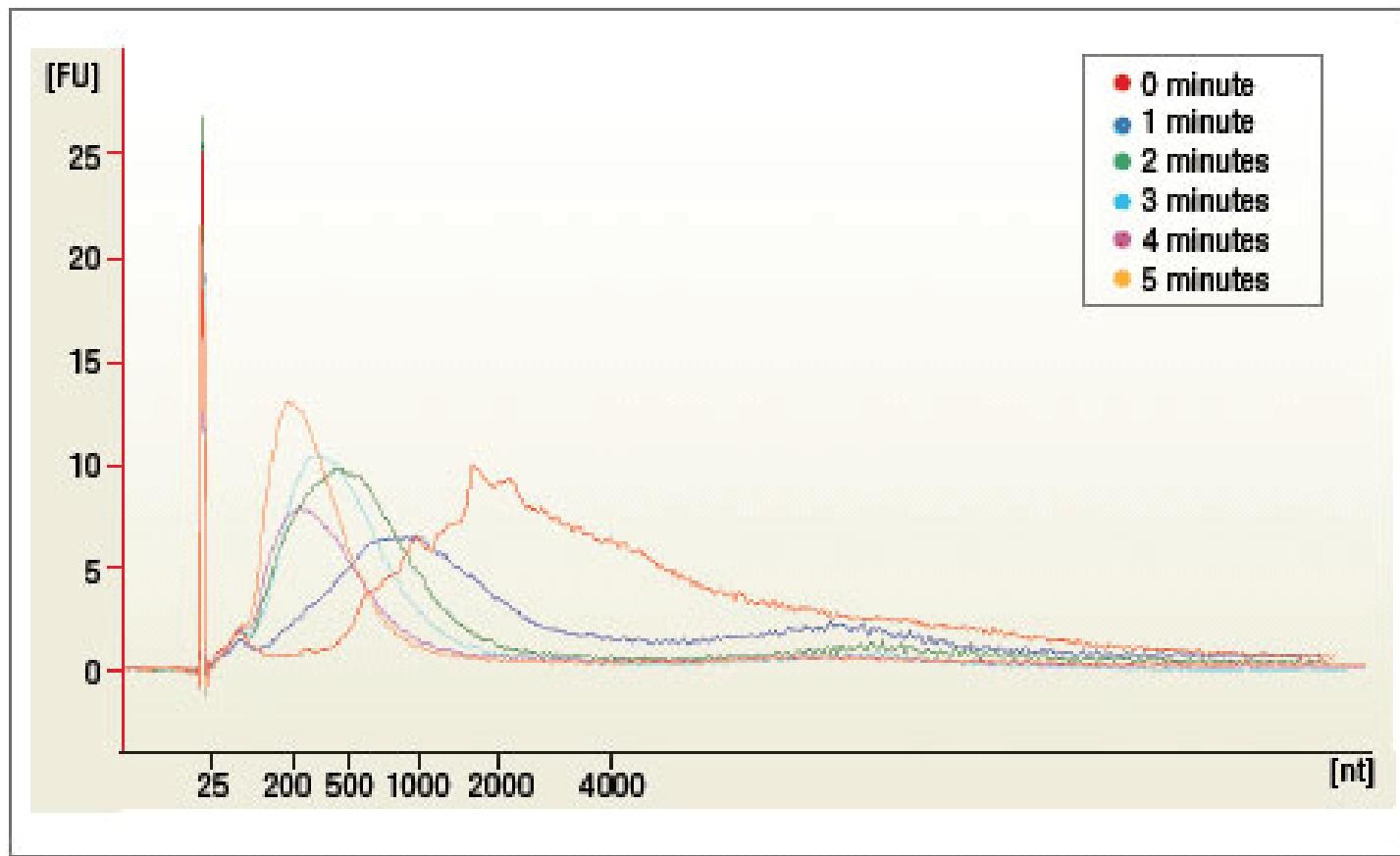
Insert size = size of insert without adapters

Mate inner distance = size of unsequenced portion
of insert after paired end sequencing

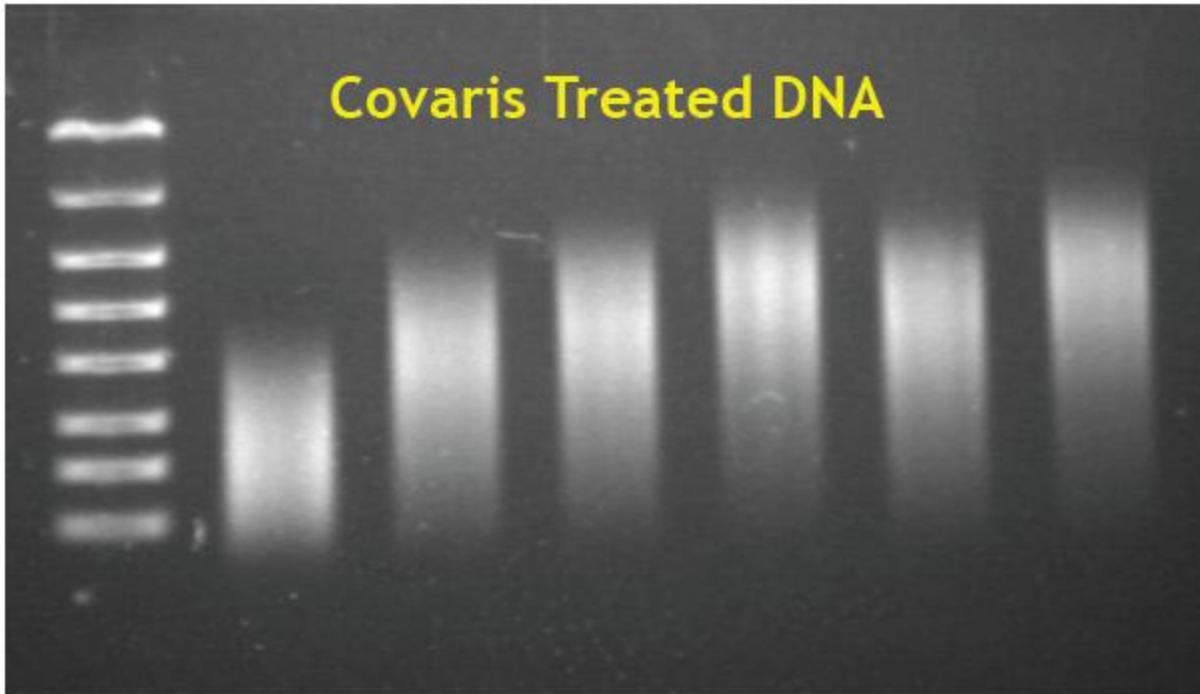
Fragment size – adapters – (2 X read length)

RNA fragmentation

Fragment RNA with Mg^{2+} and heat



cDNA fragmentation



Options for cDNA fragmentation

Nebulization

Sonicators (**Bioruptor**, Covaris)

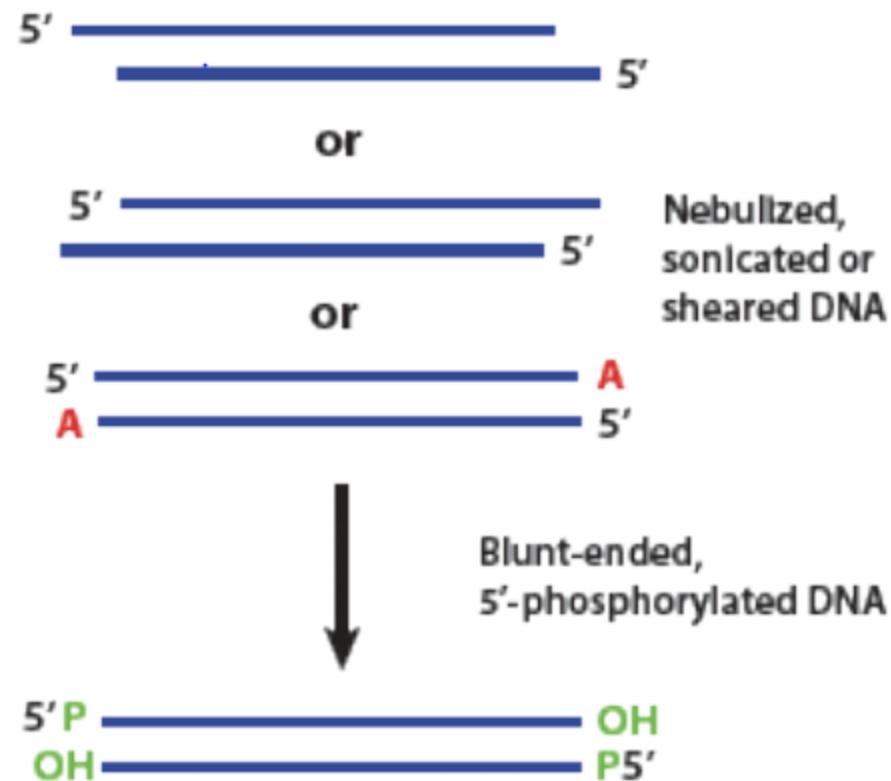
Enzymatic (can introduce bias)

Transposon (Nextera)

End Repair

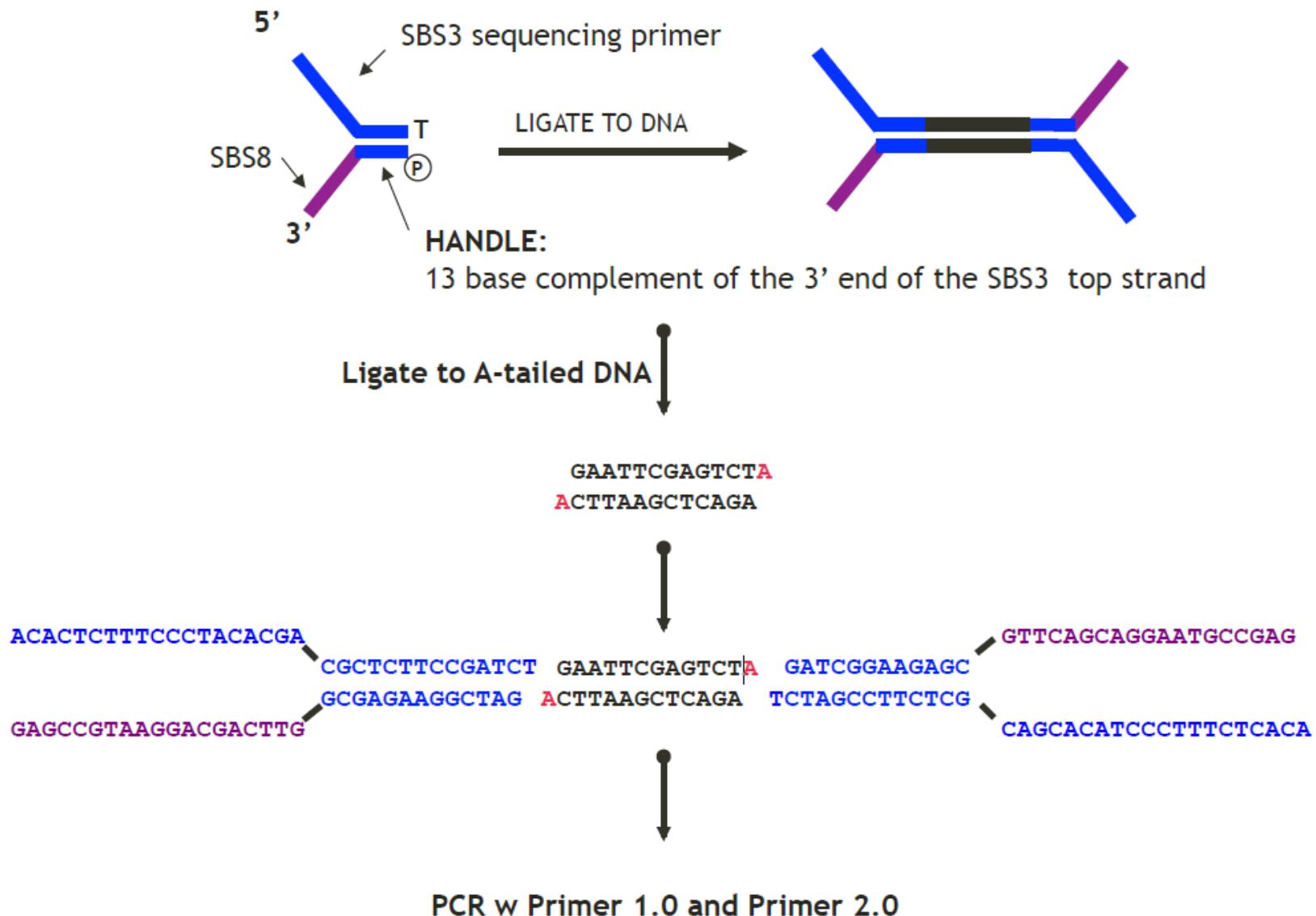
Fragmented ends
ligate poorly

Blunt the ends
(Klenow DNA
polymerase, T4 DNA
polymerase, T4 PNK)



A-tailing

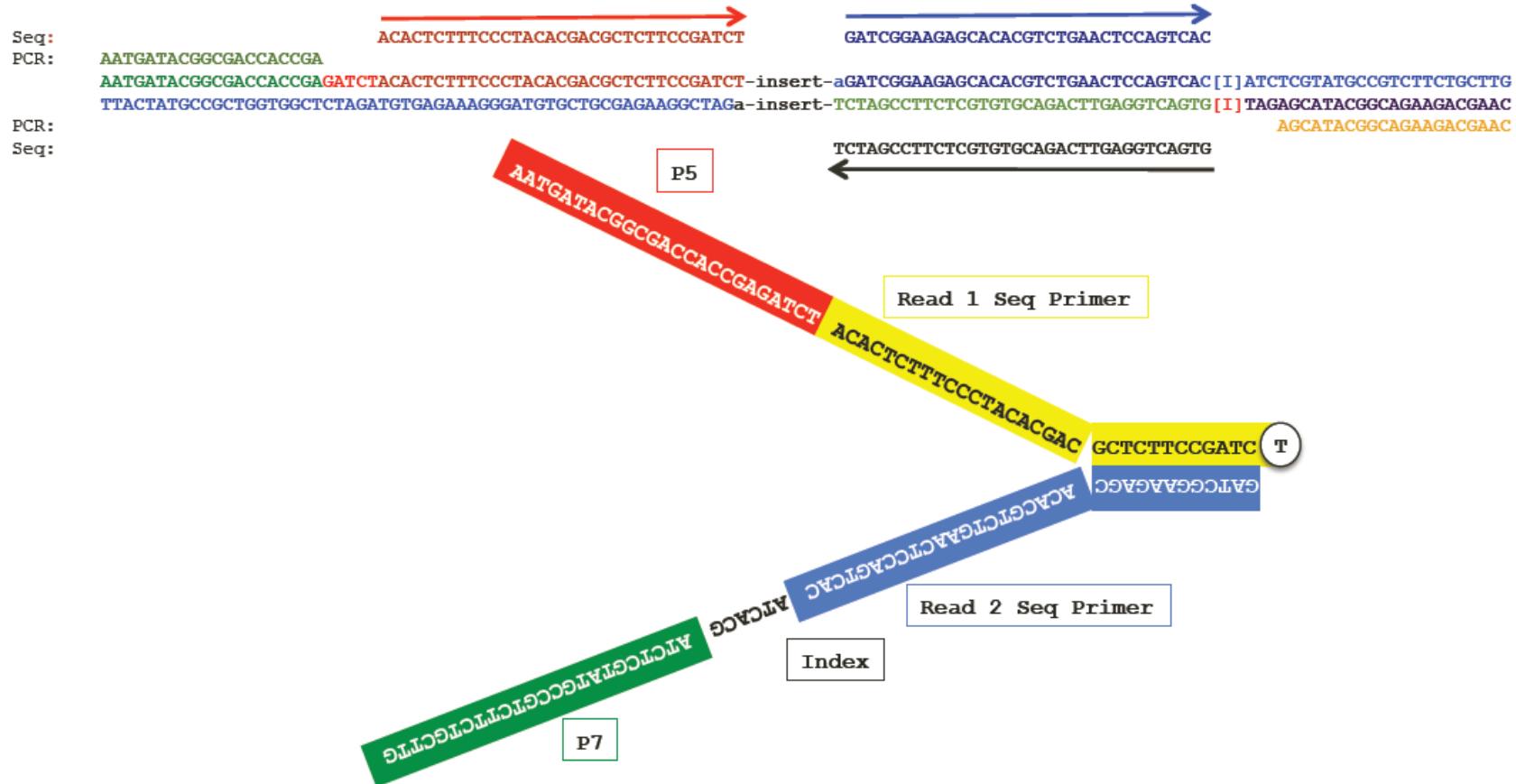
Adapter Ligation



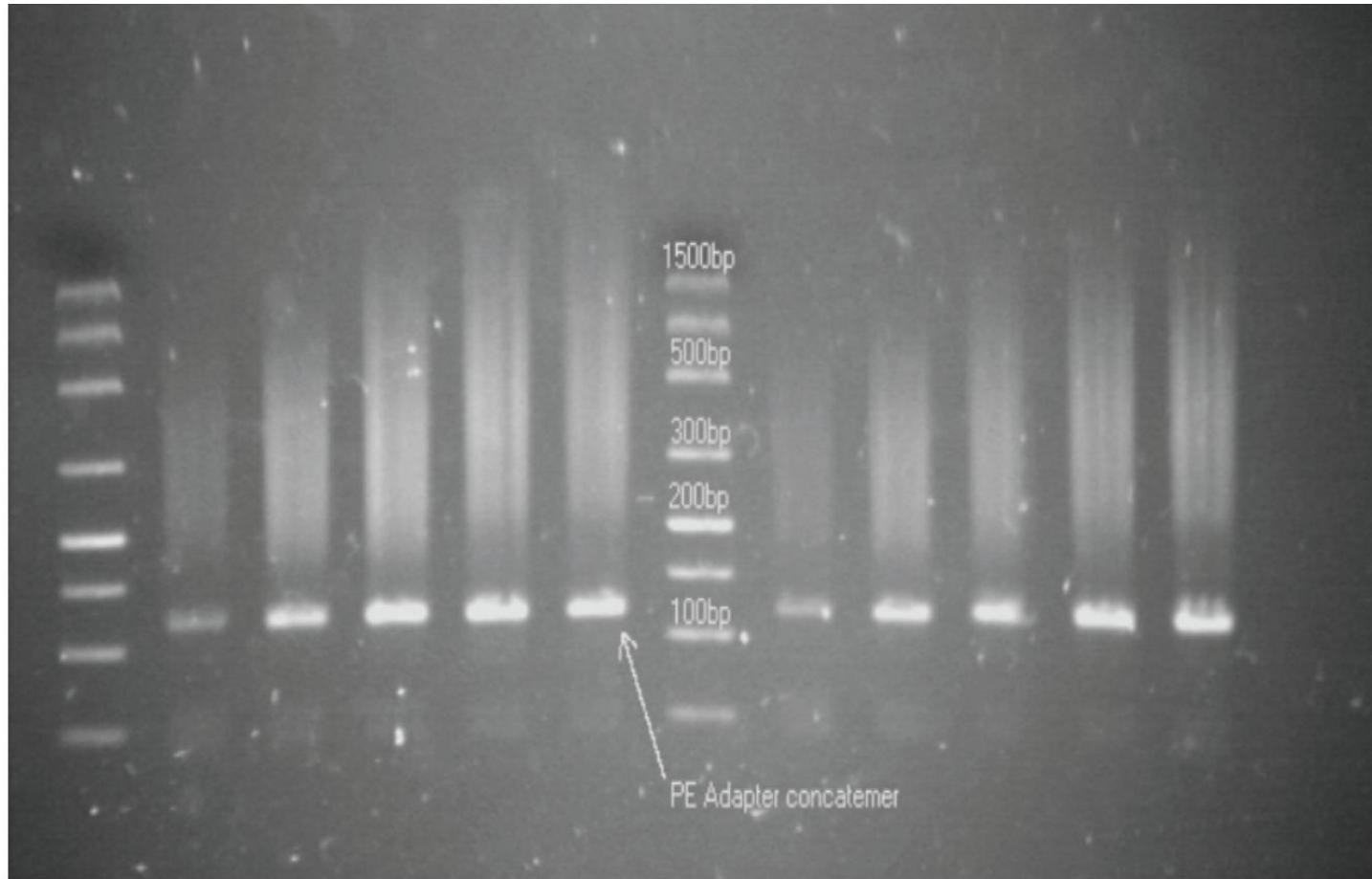
Adapter ligation

Truseq LT

- TS Univ. Adapter: AATGATAACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTTCCGATCT
- TS Index Adapter: GATCGGAAGAGCACACGTCTGAACCTCCAGTCAC [I] ATCTCGTATGCCGTCTCTGCTTG
- Primer P1: AATGATAACGGCGACCACCGA
- Primer P2: CAAGCAGAACGGCATACGA
- Multiplexing Read 1 Seq Primer: ACACTCTTCCCTACACGACGCTTCCGATCT
- Multiplexing Index Read Seq Primer: GATCGGAAGAGCACACGTCTGAACCTCCAGTCAC
- Multiplexing Read 2 Seq Primer: GTGACTGGAGTTCAGACGTGTGCTCTCCGATCT

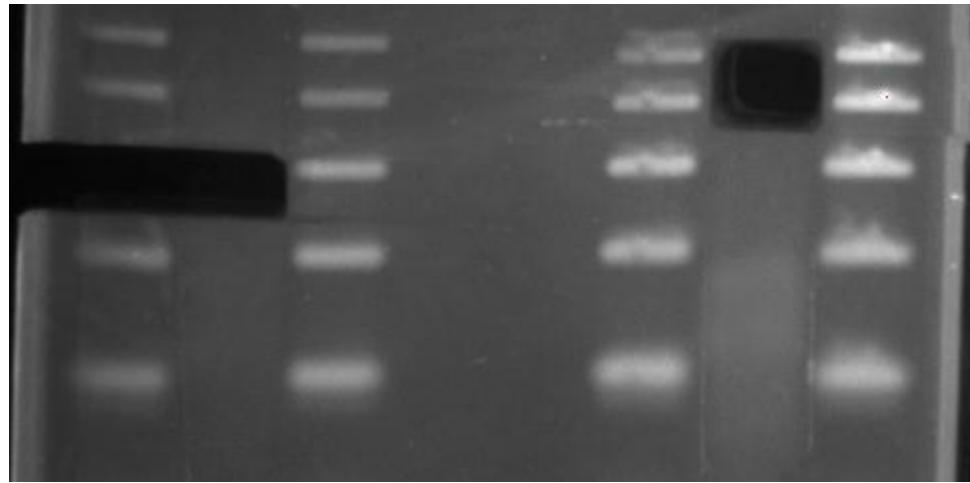


Add the ligase last!



Size Selection

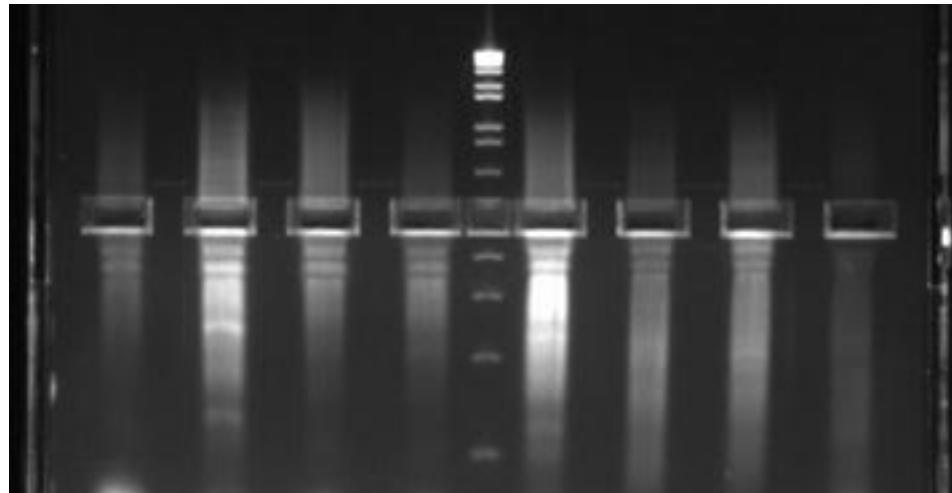
Gel electrophoresis (Agarose or Acrylamide)



Flexible
Prone to contamination
Not high throughput
High sample loss (don't heat the sample!)

Easier agarose size selection

Invitrogen E-gel System

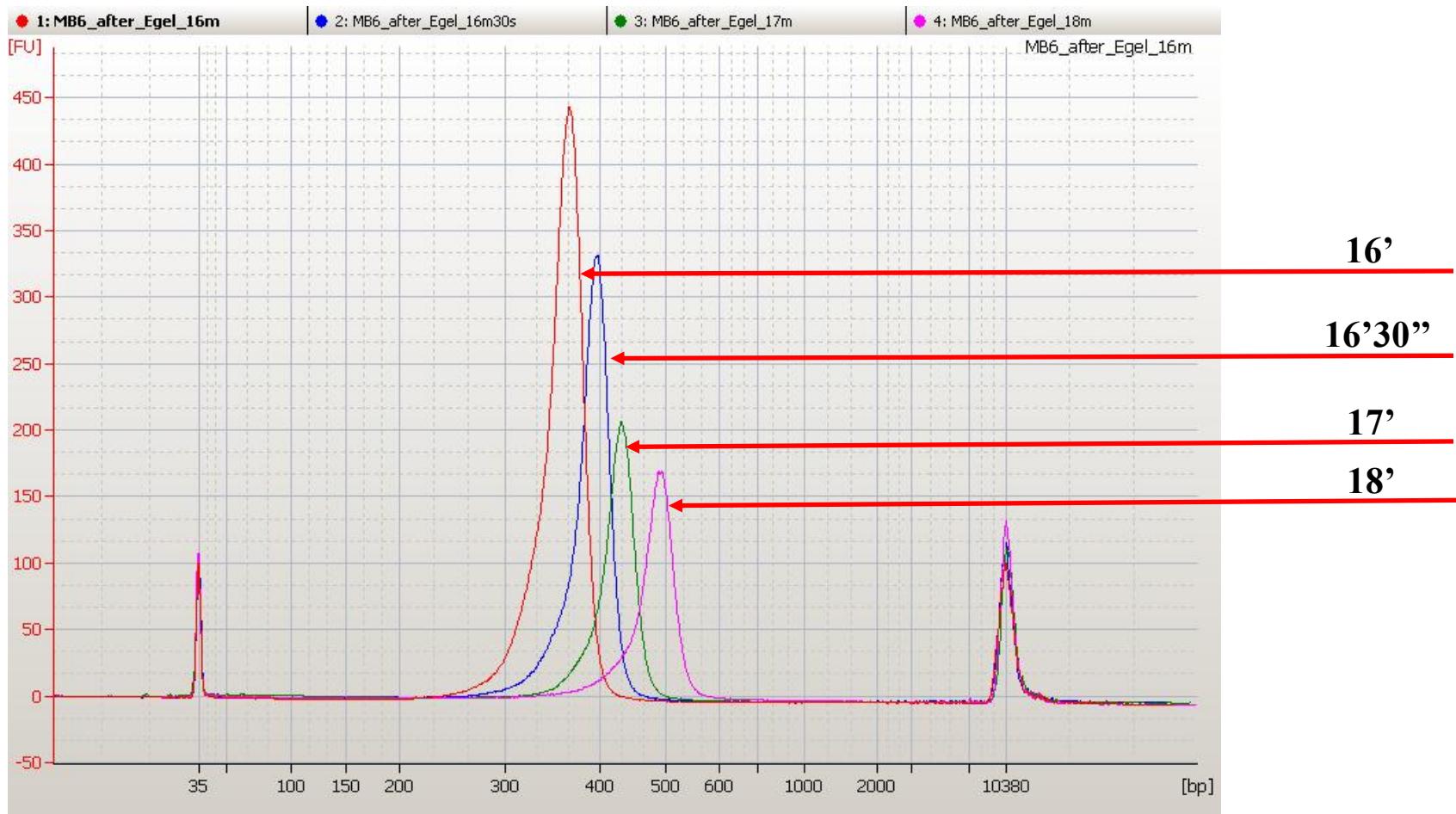


Allows repeated, regular library sizes

Provides downstream flexibility

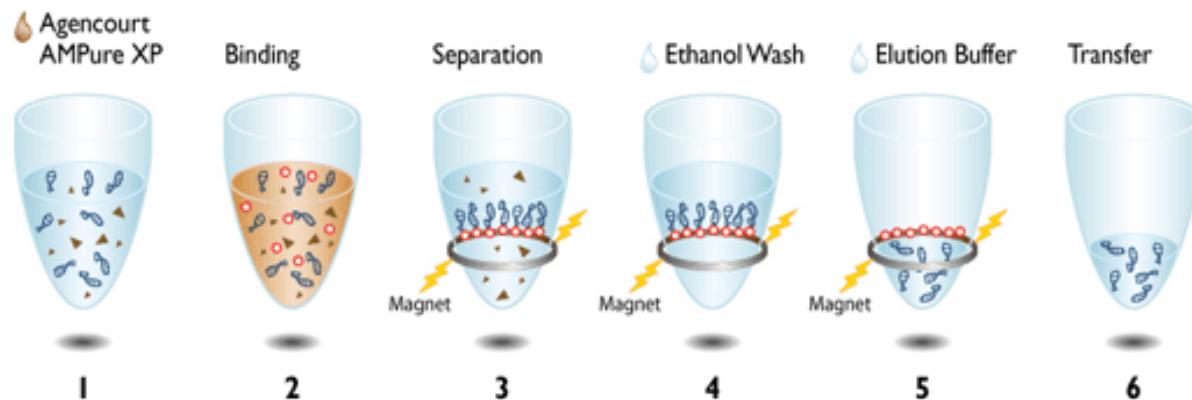
Time consuming

Invitrogen E-gel System



Ampure beads

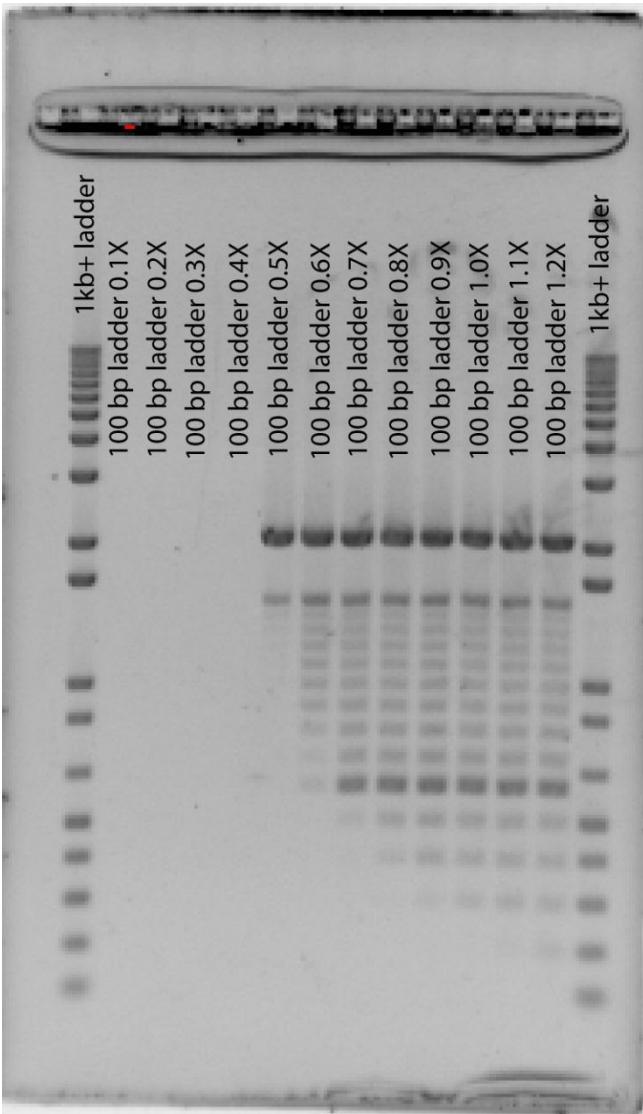
Preferred method of purification of nucleic acids for library preparation



Carboxylate-coated magnetic beads in a PEG and salt buffer

Increasing bead/DNA ratio allows smaller fragments to bind to the beads

Size Selection – Ampure beads

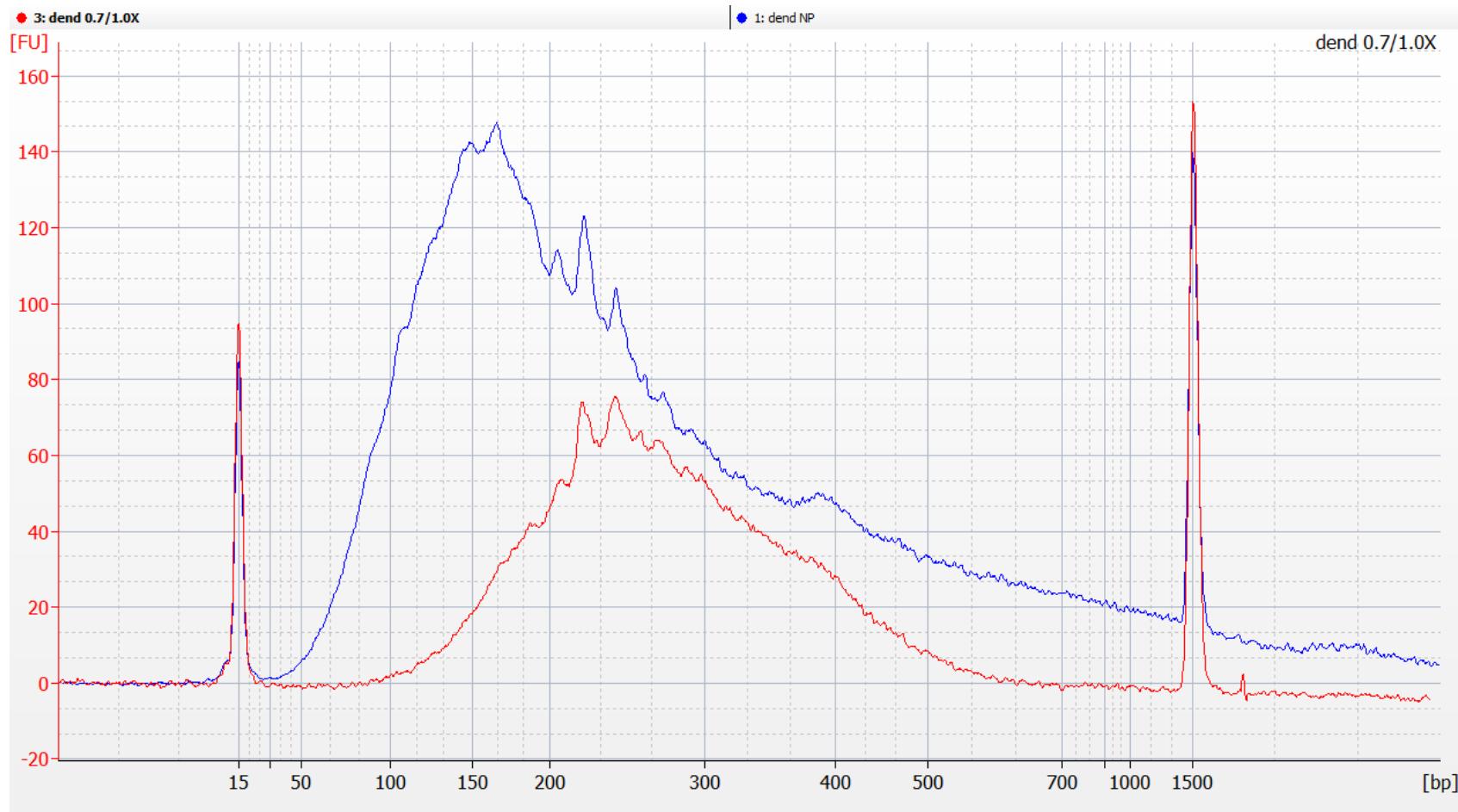


Varying the amount of beads (PEG and salt) relative to the sample allows for different sizes of nucleic acids to be retained

Best for high throughput and automation

Use 1.6–1.8X for purification

Size Selection – Ampure beads

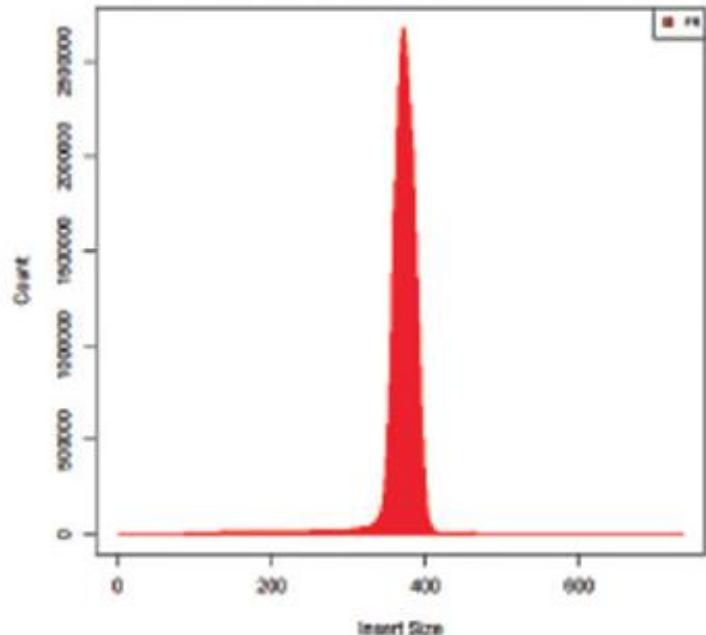
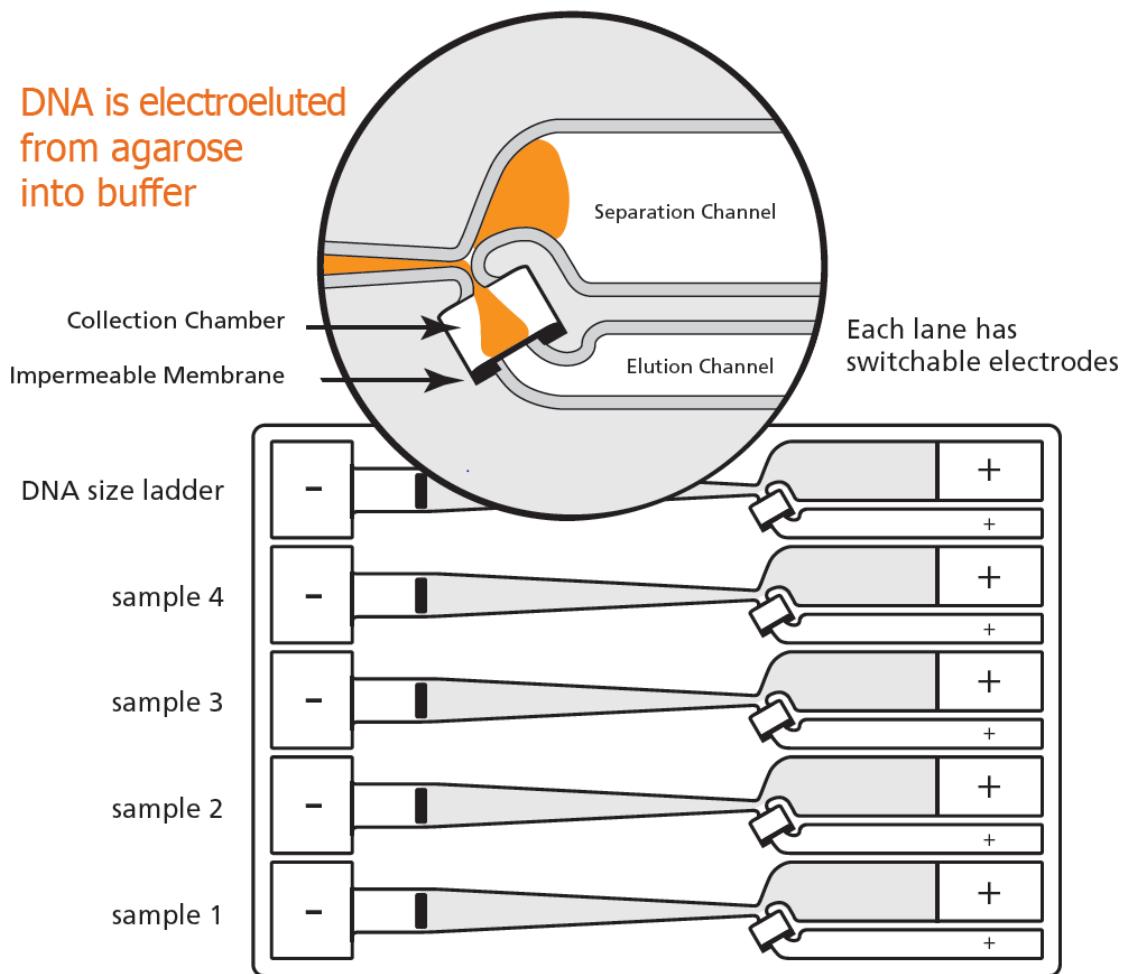


Caliper LabChip XT



Pippin Prep

DNA is electroeluted from agarose into buffer



Blue Pippin™

Collect targets between
50 bp – 50 kb



FEATURING PULSED-FIELD POWER

Planning an RNA–Seq experiment

- How will I make the cDNA?
- How will I remove rRNA, gDNA, and mitochondrial sequences?
- How will I prepare the library?
- **How many PCR cycles?**
- How much sequencing should I do?
- Should I align or assemble?
- What should I align to?
- How do I quantify expression?
- How do I test for differential expression?

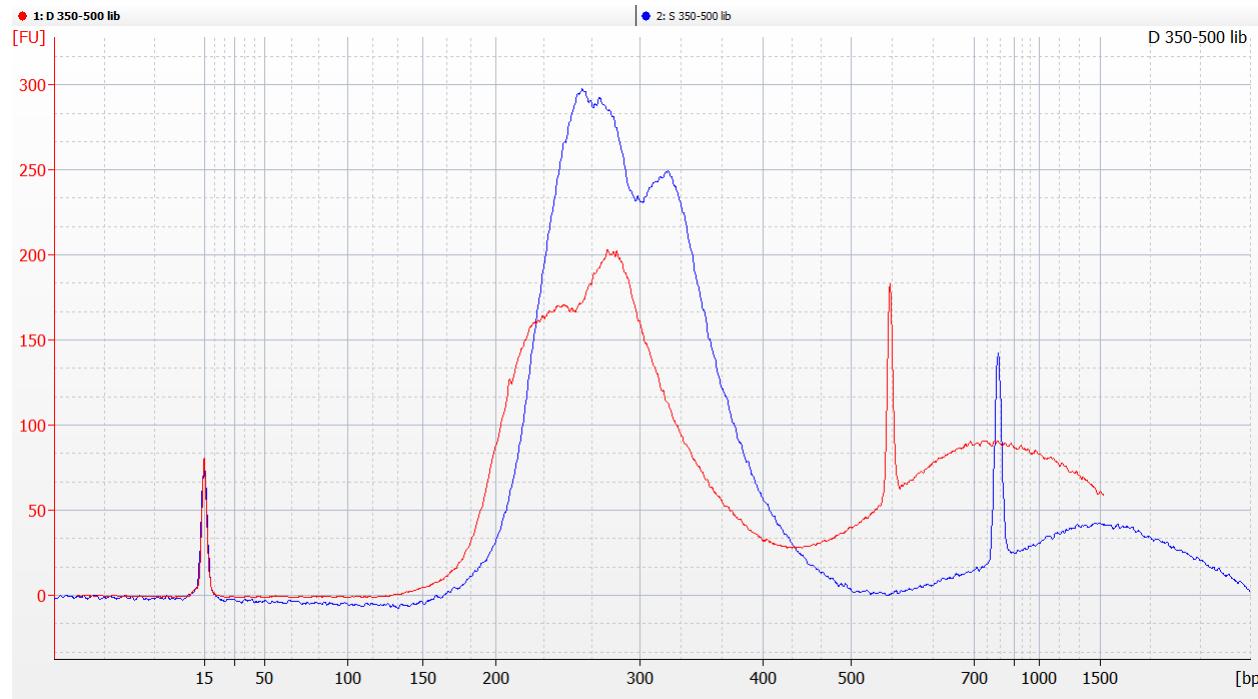
How many PCR cycles?

Goal is to minimize PCR bias

Only one cycle with large input (5ug RNA)

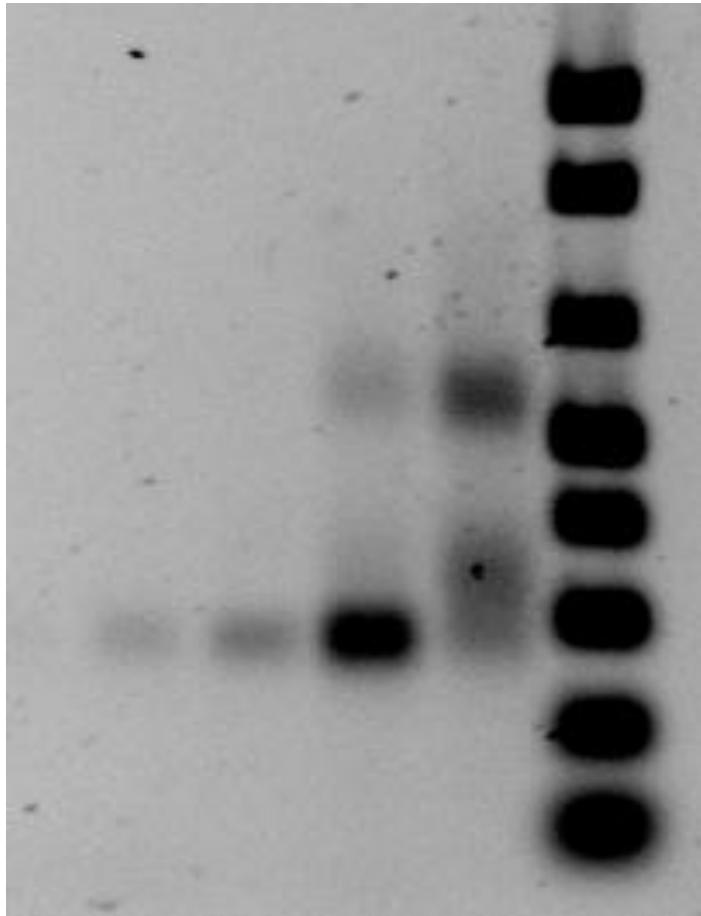
Over-amplification causes various biases

Large product from over-amplification



Optimize your PCR cycles

Manual



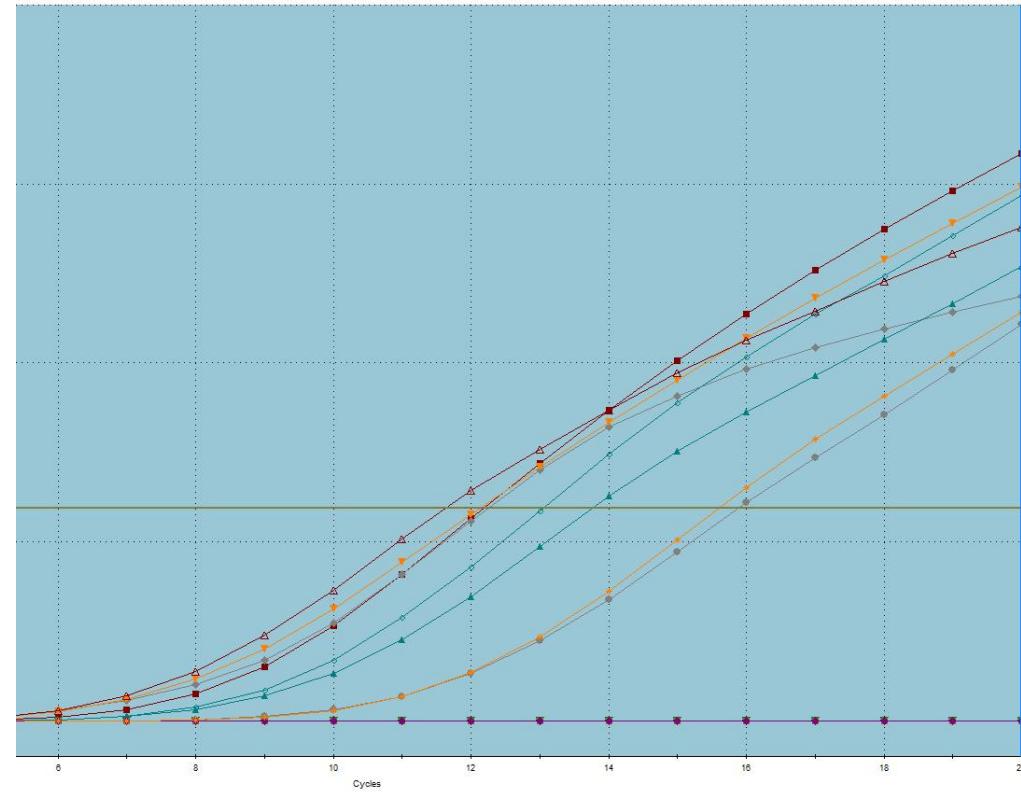
Perform 6 cycles of PCR
remove aliquot

Perform 2 cycles of PCR
remove aliquot

...

Repeat for 18 total cycles
Agarose gel

Optimize your PCR cycles



qPCR

Determine number of cycles to reach 50% amplification

For RNA-Seq, use that cycle number – 0 or 1

Set up multiple low cycle reactions, pool products

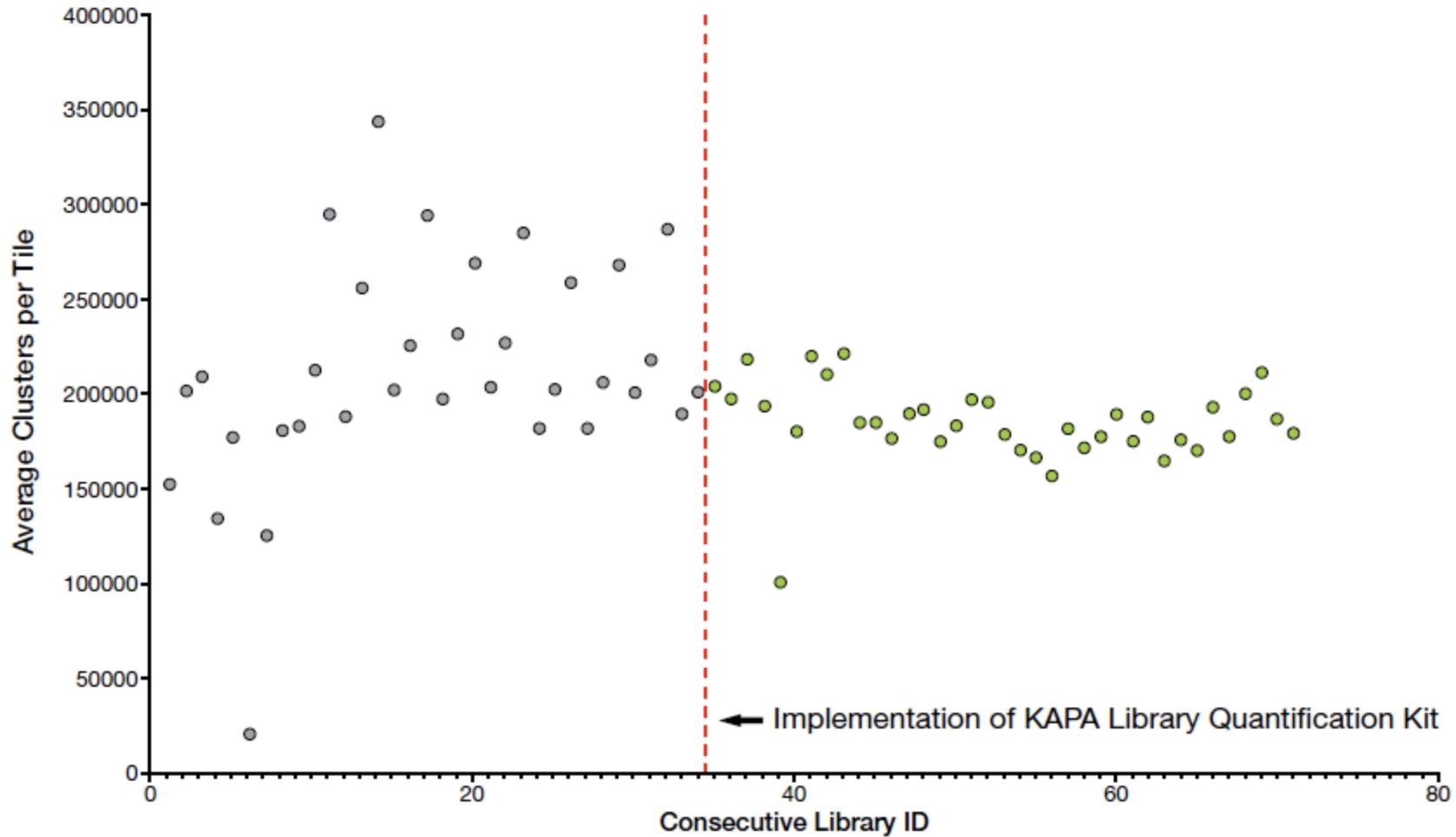
Library Quantitation

In order of increasing accuracy:

1. Nanodrop (UV spectrophotometry)
2. Bioanalyzer
3. Fluorescent Dye (PicoGreen, Qubit)
4. qPCR(KAPA Kit)

Ask your sequencing provider.

KAPA Library Quantification Kit



Minimizing technical variation

Minimize technical variation to reduce sequencing requirements

- Order all reagents needed for the experiment to ensure consistency

- Low retention, filter tip pipettes

- Low binding tubes

- Perform the library prep for all samples simultaneously or follow a randomized block design

Planning an RNA–Seq experiment

- How will I make the cDNA?
- How will I remove rRNA, gDNA, and mitochondrial sequences?
- How will I prepare the library?
- How many PCR cycles?
- **How much sequencing should I do?**
- Should I align or assemble?
- What should I align to?
- How do I quantify expression?
- How do I test for differential expression?

How much sequencing should I do?

As much as you can afford

Proper planning is crucial!

How much sequencing should I do?

What is your question?

What is your genome like?

Prokaryotic or eukaryotic?

Level of alternative splicing?

How much should you budget?

| Lane Type | Tufts / Tufts Medical Center | All Other Institutions |
|-----------------------|------------------------------|------------------------|
| Single Read, 50 Base | \$1,150 | \$1,350 |
| Single Read, 100 Base | \$1,590 | \$1,870 |
| Paired Read, 50 Base | \$1,925 | \$2,250 |
| Paired Read, 100 Base | \$2,500 | \$2,950 |

Minimum DE in a well annotated organism
with two conditions (preliminary experiment)

2 replicates per condition

30+ million reads per sample

One 1X50 HiSeq lane (\$1150)

Library prep costs (\$???)

Probably around \$3000 total

How much sequencing should I do?

Differential expression

Replicates more important than depth

Spend money on biological replicates

Sacrifice read depth and length

Human: 20–30 million reads/sample

Pool samples later for novel discovery

Must confirm with other methods

How much sequencing should I do?

Discovery

Novel genome vs. known genome

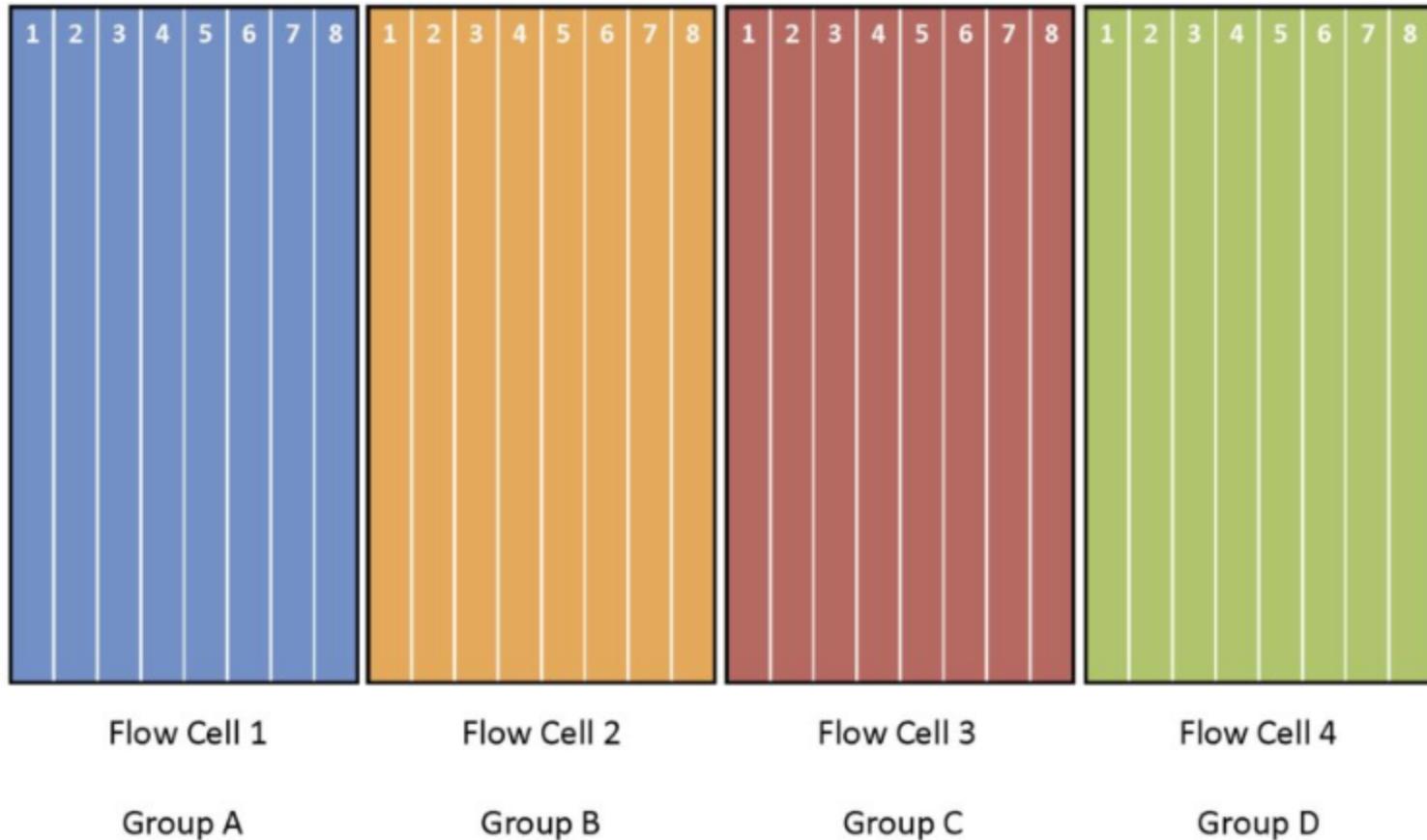
Depth and read length more important than replicates

Spend money on more, longer, paired-end reads

Sacrifice replicates (still a good idea!)

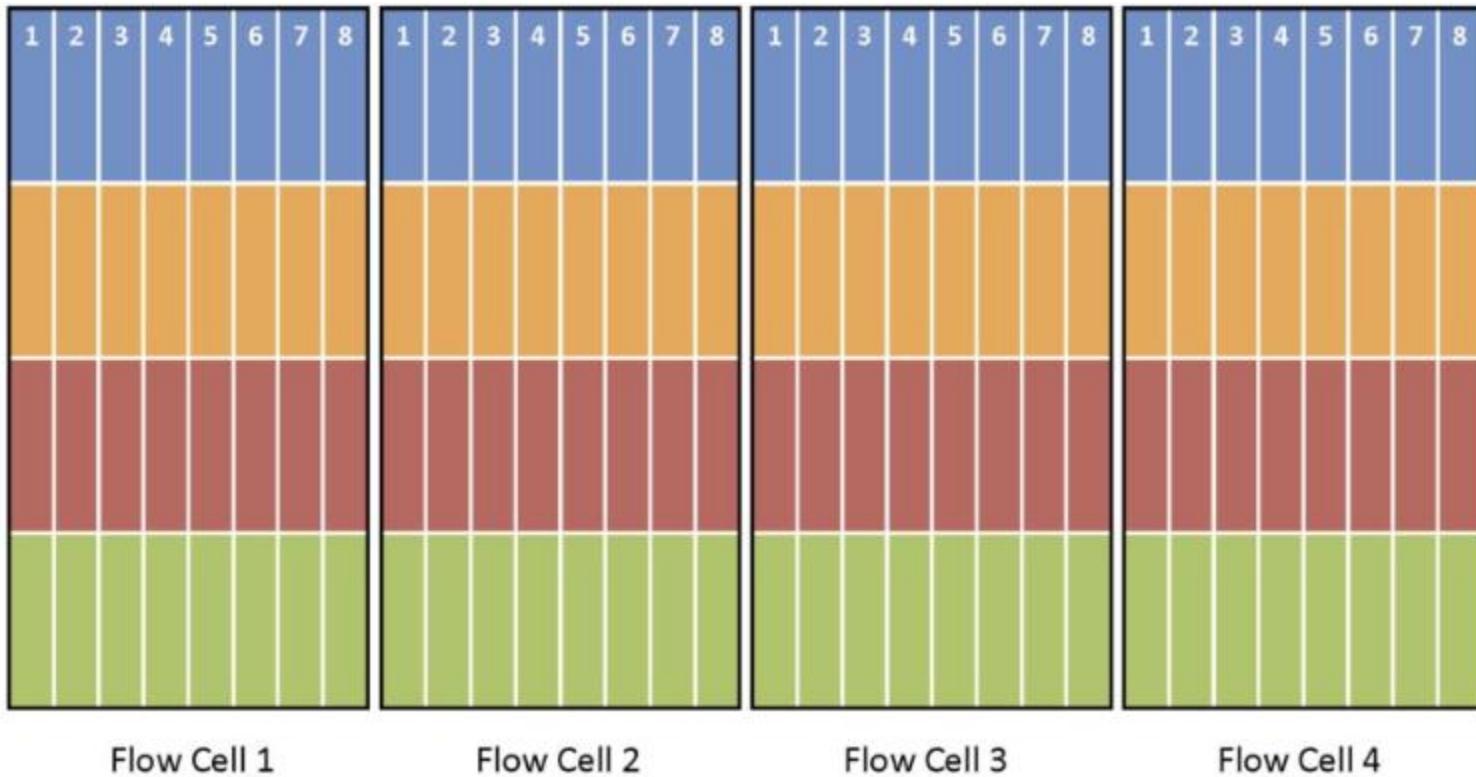
Human: 200+ million reads/sample

Poor experimental design



Cannot control for lane or flowcell effects

Multiplex for better design



Same sequencing depth
Control for lane and flowcell effects

Scotty – Power Analysis for RNA Seq Experiments

<http://euler.bc.edu/marthlab/scotty/scotty.php>

An online “tool to assist in the designing of RNA Seq experiments that have adequate power to detect differential expression at the level required to achieve experimental aims.”

Review

- Library prep involves: fragmentation, end repair, A-tailing, adapter ligation, size selection, PCR cycle optimization, library quantitation
- Do as much sequencing as you can afford.
- Make sure your sequencing is adequate to answer your question.
- Multiplexing controls for lane and flowcell effects.

Quick break

Lecture Outline

- Acknowledgements
- Computational Biology Initiative
- Course goals/outline
- Cluster test
- Sequencing technologies
- Planning an RNA-seq experiment
- Library prep
- **Design exercises**
- Some helpful resources

Questions to answer

- How will I make the cDNA?
- How will I remove rRNA, gDNA, and mitochondrial sequences?
- How will I prepare the library?
- How much sequencing should I do?

Design an RNA-Seq experiment

What alternative splicing differences are there between human and chimpanzee brains?

Design an RNA-Seq experiment

How does drug treatment affect gene expression in a pathogenic bacteria?

Design an RNA-Seq experiment

How does gene expression differ
in 200 FFPE pancreatic cancer
samples from human patients?

Design an RNA-Seq experiment

Sequence a ribosome-protected
RNA fragment from yeast.

Lecture Outline

- Acknowledgements
- Computational Biology Initiative
- Course goals/outline
- Cluster test
- Sequencing technologies
- Planning an RNA-seq experiment
- Library prep
- Design exercises
- **Some helpful resources**

Indispensable resources

1. Seqanswers

Online forum for all things NGS

2. BioStar

Question and answer site about bioinformatics

3. RNA-Seq Blog

Blog covering industry news, interesting papers, new tools

4. My NCBI

Get RNA-Seq papers sent to your email

SEQanswers

Online forum for NGS issues

Li, J.-W., et.al. (2012). SEQanswers: An open access community for collaboratively decoding genomes. Bioinformatics.

Create an account if you don't have one!

SEQanswers



You are currently viewing the SEQanswers forums as a guest, which limits your access. [Click here to register now](#), and join the discussion

| » Site Navigation | | » New Posts | | | | » Our Sponsors | |
|--|--|----------------------------------|---------|-------|-----------------|--|--|
| Title, Username, & Date | | Last Post | Replies | Views | Forum | Recent Job Postings | |
| Looking for best strategy to realign BAM files dawe | | Today 05:02 AM by dpryan | 2 | 44 | Bioinformatics | Now Hiring Senior Software Quality Engineer 01-22-2013 02:33 PM by Ingenuity Careers | |
| Annotation of sequence obtained from illumina... anishashajan | | Today 04:59 AM by GenoMax | 3 | 38 | Illumina/Solexa | Now Hiring Senior Application Scientist 01-22-2013 02:30 PM by Ingenuity Careers | |
| Service contract costs HMorrison | | Today 04:54 AM by mnelson.phd | 2 | 67 | Illumina/Solexa | Mechanical Staff Engineer 01-18-2013 03:36 PM by Pacific Biosciences | |
| Bad MiSeq Reagent Kits mnelson.phd | | Today 04:45 AM by mnelson.phd | 6 | 147 | Illumina/Solexa | Field Applications Scientist (FAS) - Menlo Park,... 01-17-2013 04:09 PM by Pacific Biosciences | |
| BWA index segmentation fault JMFA | | Today 04:41 AM by GenoMax | 4 | 74 | Bioinformatics | Next Generation Sequencing Research Associate 01-14-2013 03:07 PM by Bio Scientific | |
| NGS reads condensation bye | | Today 04:36 AM by HESmith | 3 | 83 | Bioinformatics | | |
| » SEQstandards and MINSEQE: Minimum Information about a high-throughput SeQuencing Expt | | | | | | | |
| Nov 13, 2012 - 12:20 PM - by Joann | | | | | | | |
| Introduction to the forum: a new standards page at the SEQWiki. | | | | | | | |
| Our new Wiki page hopes to be a quick and easy source of standards guidance from FGED-MINSEQE as well as various other global standards initiatives relevant to forum members with the idea of helping to put things out there sooner than later. The FGED board has chosen SEQanswers as the most appropriate location in the entire webverse to host feedback opportunities on their consensus format for ultra high throughput sequencing experiments. The forum is a great place to conduct discussions, and already has been active in exploring terminology issues and more. This thread would be a place to start collecting these discussions, plus feedback, inquiries and input from our forum members and also to refer students, especially, to the new Wiki page material on formal standards guidelines for sequencing and publishing sequence data. | | | | | | | |
| 5 Replies 1,458 Views | | | | | | | |

Forums

Wiki – Find RNA-Seq software

Instrument Map

Want to keep up with what's new?



Twitter: @agbt

Hashtag: #AGBT

RNA-Seq Blog

Other blogs and resources on course website
<http://sites.tufts.edu/cbi/resources/rna-seq-course/>

Questions? Minute cards!

- Acknowledgements
- Computational Biology Initiative
- Course goals/outline
- Cluster test
- Sequencing technologies
- Planning an RNA-seq experiment
- Library prep
- Design exercises
- Some helpful resources